1963

Methods for phonemic recognition in speech processing.

Hollabaugh, Jon Dale.

Monterey, California: U.S. Naval Postgraduate School

http://hdl.handle.net/10945/12661

METHODS FOR PHONEMIC RECOGNITION IN SPEECH PROCESSING

JON DALE HOLLABAUGH

23

METHODS FOR PHONEMIC RECOGNITION IN SPEECH PROCESSING

* * * * * *

Jon Dale Hollabaugh

METHODS FOR PHONEMIC RECOGNITION IN SPEECH PROCESSING

by

Jon Dale Hollabaugh

Captain, United States Marine Corps

This work is accepted as fulfilling
the thesis requirements for the degree of
MASTER OF SCIENCE
IN
ENGINEERING ELECTRONICS
from the
United States Naval Postgraduate School

ABSTRACT

Speech is one of the most inefficient methods of communication. Therefore, there has been a continuing effort to devise means to reduce the redundancy, that is, compress the bandwidth required for speech communication. For broad tactical military use the author considers the amount of compression, intelligibility and quality to be factors of prime importance, while factors such as speaker recognition and naturalness are of secondary importance. The speech compression methods receiving the most emphasis today are described and their major discrepancies indicated. Generally, their deficiencies arise because the present systems do not rely on the fact that the electrical representation of speech is a particular signal, not just any electrical signal whose frequency components lie in the audio band.

With present day speech compression systems in mind, an analysis of the "method of distinctive features" as proposed by Jakobson, Fant, and Halle of MIT is offered. This method for achieving reliable speech recognition at the phonemic level is in a partial stage of development. The measurements required to extract six of the ten features are tabulated and procedures for reducing the remaining four features are outlined. Sonagraphic data is given in support of the method.

Instrumentation and flow charts for digital computer investigation of the process are detailed. The proposed acoustical input device will facilitate computer studies of

ABSTRACT
(Continued)

speech processes.

The author would like to thank Dr. S. E. Gerber and Mr. Fausto Poza of Hughes Aircraft Company for their assistance and encouragement.

The writer also wishes to express his appreciation to Professor Mitchell L. Cotton of the Digital Control Laboratory, U. S. Naval Postgraduate School, for his guidance toward this interesting facet of communications.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS
(Continued)

1. Introduction

The military requirements for communication facilities of all types are increasing constantly. This is largely due to expanding missile and satellite applications. In conjunction with this rapid increase in military requirements, electronic engineers have made substantial progress in the field of digital communications in an effort toward more efficient utilization of the available bandwidth. Since no military commander will relinquish his voice net, and realizing the redundancy inherent in speech, it seems logical to apply digital methods to speech processing. In addition, a digitized speech communication link may be made "secure" rather easily.

For broad tactical military use, we are primarily interested in bandwidth compression, intelligibility and quality. Factors such as speaker naturalness and recognition are of little concern except in special isolated cases.[1] However, this concept has not been followed in the development of "vocoder" and pulse code modulation techniques. Generally, these methods consider any electrical signal whose frequency components lie between zero and ten kilocycles as "speech". This generalization is the major contributing factor to the poor compression ratio inherent in digital speech systems today. True speech has in fact, certain distinctive features which when measured, uniquely define phonetically what was

---

[1] See Appendix A for definition.

said. This idea was offered by a group of linguists and phoneticians at the Massachusetts Institute of Technology and it is this idea which forms the basis of this thesis /1/. There have been similar theories but this "method of distinctive features" is unique in that while linguistically sound, the binary processing indicated lends itself neatly to reduction to electronic hardware. This method, with respect to the state-of-the-art, is considered by many as the most auspicious approach to the problem of speech recognition. The ultimate goal is a machine which accepts a speech wave at its input and generates a sequence of phonetic symbols at its output; as a synthesizer, it accepts a sequence of symbols at its input and generates a speech wave /2/. A speech recognizer with this capability could transmit phonemic information at rates less than fifty bits per second (for English).

As an initial step toward a speech research program at the U. S. Naval Postgraduate School, the complete design of an acoustical input device, compatible to the CDC 1604 via the CDC 160, is given. For a nominal cost, this device will provide a means for the investigation of the "method of distinctive features" as well as any theory related to speech recognition and/or communication. Speech samples will be transmitted to the computer memory to form a time-amplitude-frequency spectrum therein, which may be operated on in any manner desired.

Programs for the recognition of the distinctive features will be flow-graphed to provide the basis for future investigations.

2. Speech Production and Transmission

For many years, communications engineers have recognized the importance of trying to match the transmission channel to the source of information. The classical example of non-equivalence has long been conventional telephony. Most of the voice systems in use today operate on the same fundamental principle implemented for the first time by Alexander Graham Bell about 86 years ago. The principle, of course, is facsimile reproduction of acoustic waveforms. Consequently, most conventional voice channels are capable of transmitting information at about 30,000 bits per second. However, the information rate associated with the phonemically-transcribed equivalent of speech is about 50 bits per second /3/. In addition, it has been determined that the rate at which a human is able to assimilate information is of the same order, i.e. 50 bits per second /4/. It is true that these figures do not include man's ability to extract prosodic information from speech and if this is taken into account, the information rate is somewhere around 1000 bits per second. So there is still a great disparity between the information rate of the human source and the channel capacity of the conventional telephone circuit.

The usual approach to matching the speech transmission channel to the information source is:

(1) Determine the constraints which are characteristic of the production and perception of speech;

(2) Incorporate these constraints into the transmission system;

(3) Determine an efficient means for transmitting the parameters of the constraints.

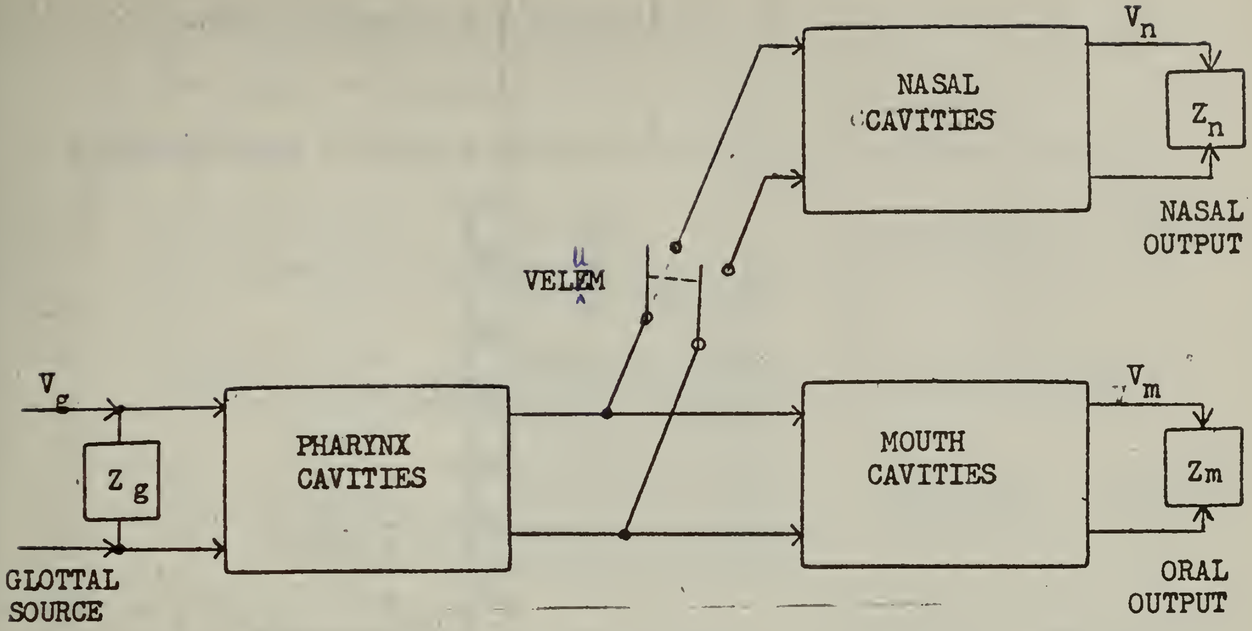
Systems built according to this outline are commonly called analysis-synthesis systems.

With respect to (1) above, the vocal tract may be viewed as an acoustic tube of varying cross-sectional area which is excited by a periodic or noise source. The discrete nature of the source of vocal excitation suggests a system-function analysis in which the effects due to the excitation and the passive system may be distinguished (Figure 2-1). The top sketch represents the production of sounds when excited by the vocal cords. Sounds so produced are vowel-like, such as /a/, /i/, /l̩/, and /v/ and are commonly called "voiced" in the literature. The glottis[1] is represented by a source of acoustic volume velocity V_g, with an internal impedance Z_g. The radiation impedances at the mouth and nostrils and volume velocities through these impedances are denoted by Z_m and Z_n, V_m and V_n, respectively.

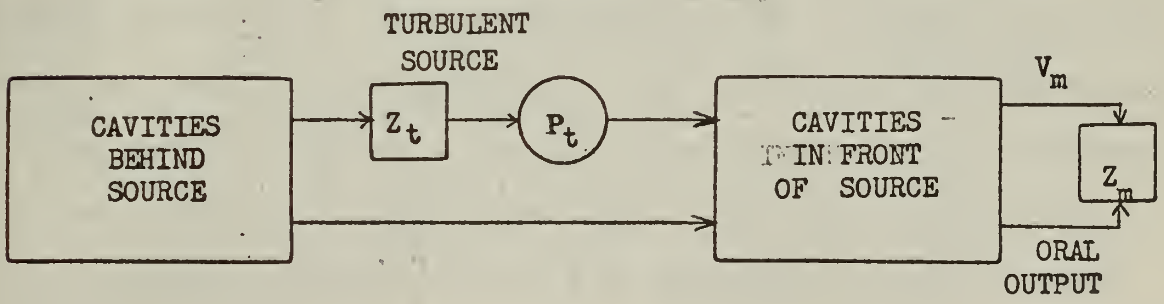
The lower sketch represents the production of "unvoiced" sounds, such as /p/, /s/, /t/, and /ʃ/. In this case, the sound source is located further along in the tract and is represented by a series pressure source, P_t, with inherent impedance, Z_t. For these sounds, radiation occurs mainly from the mouth.

---

[1] See Appendix A for definition.



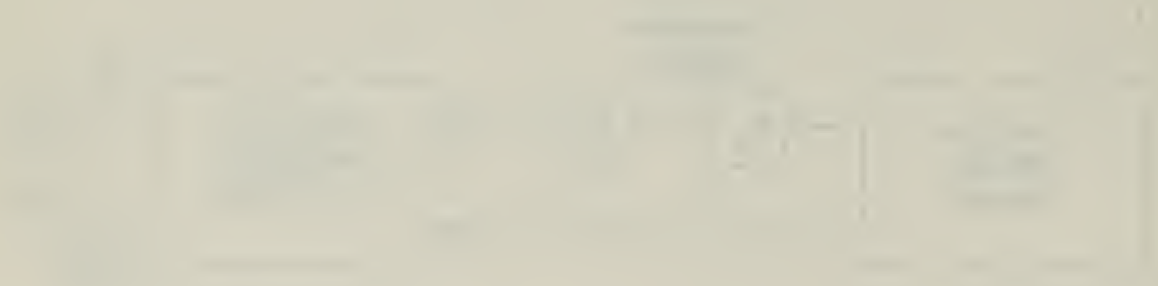
GLOTTAL EXCITATION



NOISE EXCITATION

Figure 2-1. Block diagram representation of the vocal mechanism.

[illegible]



[illegible]

Let us write the transfer functions of the oral cavity for the cases of nonnasalized voiced, nasalized voiced, and nonnasalized unvoiced sounds in Laplace transform notation.

$$(1) \quad \frac{V_m(s)}{V_g(s)} = \frac{\prod_{\infty} s \cdot s_k^*}{\prod_{\infty} (s-s_k)\cdot(s-s_k^*)} \quad ;\ \text{nonnasal, voiced}$$

$$(2) \quad \frac{V_m(s)}{V_g(s)} = \frac{\prod_{\infty} (s-s_m)\cdot(s-s_m^*)}{\prod_{\infty} (s-s_n)\cdot(s-s_n^*)} \quad ;\ \text{nasal, voiced}$$

$$(3) \quad \frac{V_m(s)}{P_t(s)} = \frac{\prod_{\infty} (s-s_j)\cdot(s-s_j^*)}{\prod_{\infty} (s-s_k)\cdot(s-s_k^*)} \quad ;\ \text{nonnasal, unvoiced}$$

For nonnasalized voiced sounds, the variant features of the transfer function are specified by the poles of the vocal tract. Here, the normal modes of the tract are in general manifested in the acoustic output as nearly constant bandwidth spectral maxima which are called "formants".[1]

However, equation (2) indicates that for nasalized voiced sounds, the variant features of the transfer function involve zeros as well as poles. The zeros are anti-resonances of the vocal tract and occur when the driving point impedance of the nasal tract approaches zero. The pole configuration is determined in this case by the constants of the coupled systems.

Finally, for nonnasalized unvoiced sounds, the zeros of equation (3) are introduced by virtue of the location of the source. The oral output will have zeros at frequencies where

[1] See Appendix A for definition.

the impedance looking from the source back toward the glottis is infinite.

The motions and permissible values of the poles and zeros of the vocal transfer function reflect the physical motions and physiological constraints of the speech mechanism. In addition, we know that even under the most favorable circumstances, man cannot detect errors of less than three percent in data specifying a formant frequency. Therefore, we needn't try to measure formant frequency data any more accurately than that. Also, it has been shown that only the first three formant frequencies need be determined for use in the synthesis of good quality speech.

From this system function point of view, speech communication might be reduced to a signaling of data specifying the vocal excitation function and the vocal transfer function. All analysis-synthesis schemes follow this basic idea; however, none of them are the precise analog of equations (1), (2), and (3) /5/.

Having developed the logical basis for the analysis-synthesis scheme of speech compression, let us simply state the central problems:

(1) Determine the most suitable form in which to code the speech signal, such that the channel capacity required for transmission of the coded signal is as low as possible;

(2) Develop instrumentation to extract the data in

(1) with minimum error, and to receive the coded data and synthesize acceptable speech.

Implicit in this discussion is the requirement that the output speech from the synthesizer meet some criterion of fidelity, as determined by suitable psycho-acoustic tests /6/.

The next four sections of this thesis explain the theory and circuitry of the four major classes of speech compression systems. They are:

(1) Fixed channel vocoders, based on frequency domain analysis;

(2) Correlation vocoders, based on time domain analysis;

(3) Formant tracking vocoders, based on parametric analysis;

(4) Various hybrid types, based on some of the good characteristics of the previous three classes and other original ideas.

Most of techniques are designed for the frequency range zero to four kilocycles per second. There has been some work done on so-called "Hi-Fi" vocoders in which the frequency range zero to ten kilocycles is compressed into the telephone band, but this result is of little military value /7/.

3. Fixed Channel Vocoder

The first piece of equipment built on the analysis-synthesis concept was constructed by Homer Dudley in 1939.[1] The "Vocoder", or voice coder, envolved due to the results of intensive studies of the physical reproduction of speech conducted by Harvey Fletcher in the mid 1920's and by Dudley in the 1930's /8, 9/. Their studies resulted in a better understanding of the part played by the vocal cords, lips, tongue, teeth, nasal passages and lungs in the production of speech. In addition, they pointed out wherein speech is redundant and which parts are significant in conveying intelligibility. In achieving a compression ratio of 10:1, the vocoder incorporates one constraint of production and one of perception, namely, it recognizes that speech may be voiced or unvoiced, and that intelligibility may be maintained to a large extent, by preserving the short-time amplitude spectrum.

A block diagram of the classical fixed-channel vocoder is shown in Figure 3-1. A set of contiguous bandpass filters, each with its own rectifier and low-pass filter, produce values of the short-time spectrum at discrete frequency points. To a first approximation, the magnitude of the vocal transfer function is thereby produced. A separate channel, the pitch extractor channel, is designed to develop a voltage proportional to the fundamental vocal frequency of the voiced sounds and this same voltage indicates voiced-voiceless

---

[1] Dudley built the "Voder", or Voice Demonstrator, in 1938, but this was purely a synthesis device.

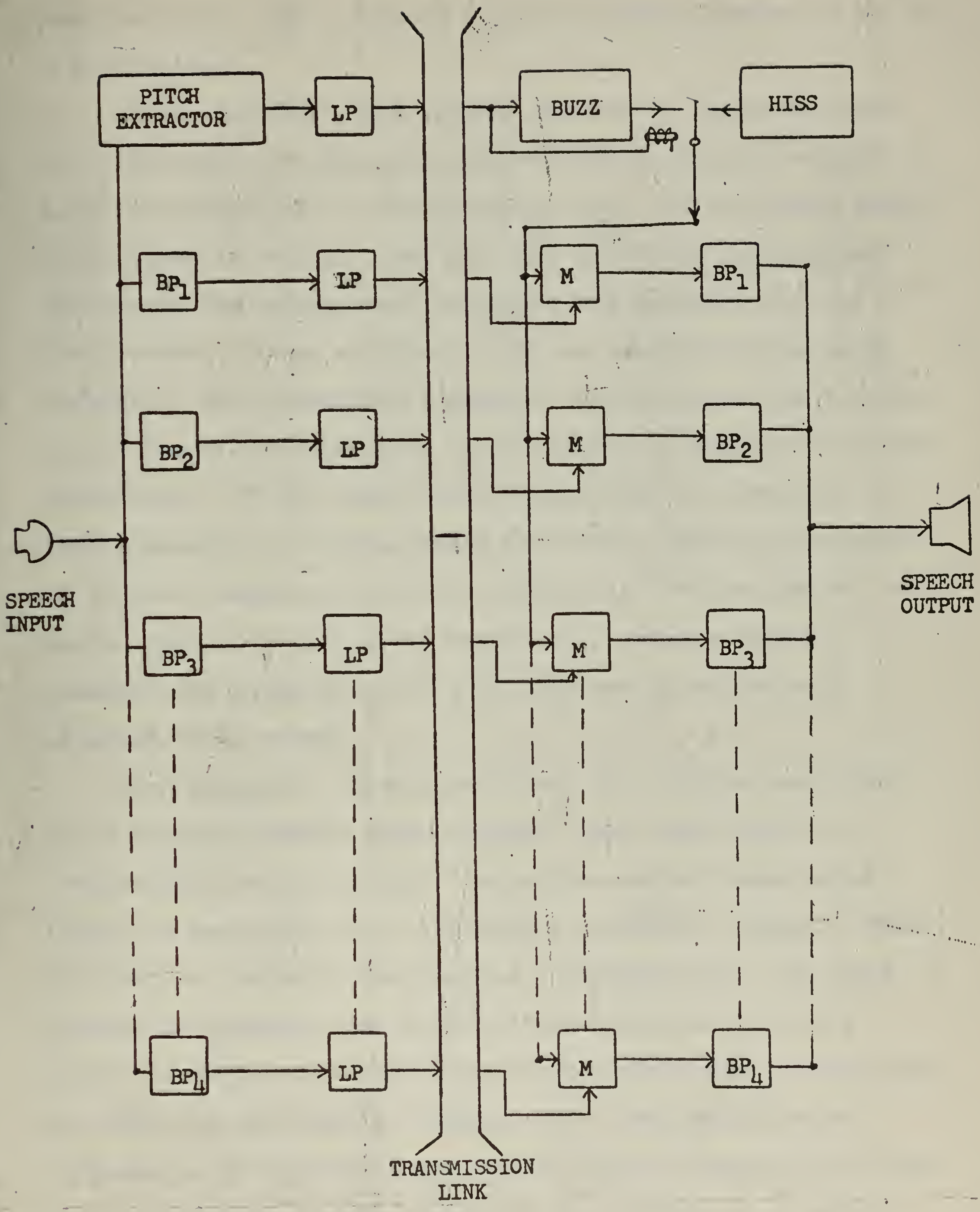


Figure 3-1. Block diagram of a fixed-channel vocoder.

distinctions. This channel is also low-pass filtered prior to transmission.

At the receiver, the inverse process is performed with the exception that the pitch information is used to control a buzz-hiss generator in the following way. If the sound being transmitted is voiced, then the buzz generator is energized generating the fundamental frequency and harmonics of the vocal cords. These signals are fed as carriers to balanced modulators on the output of each of the filters. The signals out of the filters modulate the carriers (if present) in these modulators. If the sound being transmitted is unvoiced, the hiss generator is activated and produces a noise-like signal of constant amplitude which is applied in the same manner as stated above for the voiced condition. It is common to consider any sound which is a combination of voiced and unvoiced to be voiced.

Each frequency channel requires about 20 cps bandwidth and a signal-to-noise ratio somewhat less than that of a conventional voice circuit. The pitch channel needs about twice the bandwidth of an individual frequency channel. With 18 spectrum channels spanning the telephone band, and using digital techniques, the vocoder transmitter can transmit highly intelligible speech at an information rate of the order of 2000 bits per second. However, the compression ratio obtained is not as high as we would like for tactical military use. In addition, since there are 19 channels to be multi-

plexed, it is very likely that the modulating method employed will add considerably to the bandwidth required for transmission /10/.

There are two major problem areas to be considered in the design of a vocoder transmitter, i. e., the filter bank and the pitch extractor channel. With respect to the number of filters, a trade-off must be made between such factors as high intelligibility, naturalness, and speaker recognition (indicating a large number of filters) versus desired bandwidth compression (the larger the amount of compression desired, the smaller the number of filters). Research has shown that 16 to 18 filters will give high intelligibility and fair naturalness and speaker recognition, while giving a compression ratio on the order of 8:1 /11/. The filters are usually closely spaced at the low frequencies, in an effort to obtain good formant frequency information, and more widely spaced at the higher frequencies as the energy is considerably lower in these bands. However, the wider the bandwidth, the greater the averaging of the harmonics and consequently, less information is retained. On the other hand, if the bandwidths are made too narrow, the formants may "glide" from one channel to the next, increasing the formant tracking problem. The band-pass characteristic of each filter is typically flat in the pass-band with a 40-80 decibels per octave slope on the skirts. The skirts of adjacent filters should intersect at the ten db down point at the edge frequencies. The low-pass

filter on each channel is of the same quality as the bandpass filters, with the cut-off frequency being typically 20-25 cycles per second. The output of each low-pass filter is sampled at the Nyquist rate[1], but frequently a slower rate is used as there is little change in intelligibility for slightly slower rates /12/. With respect to quantization in amplitude, eight amplitude levels are considered very conservative and some systems use just one bit.

The pitch extraction problem and the related buzz-hiss switching problem are common to most analysis-synthesis speech systems. In any pitch extraction system the fundamental frequency of the voice must be physically present in the input or be reconstituted from the remaining harmonics. A typical pitch extraction circuit is shown in Figure 3-2. The fundamental must be emphasized sufficiently so that all the axis crossings prior to clipping, are due to the fundamental and not higher harmonics. Some sort of fundamental filter is needed to accomplish this emphasis. In the early vocoders, a male-female switch was used to select the fundamental pitch range for men or women, and in later types, tracking filters were used. These filters were made to start at low frequency and increase in frequency until the fundamental was emphasized and the second harmonic suppressed /10/.

When a sound is voiced, the pitch extraction channel is energized and information concerning the location of the

---

[1] See Appendix A for definition.

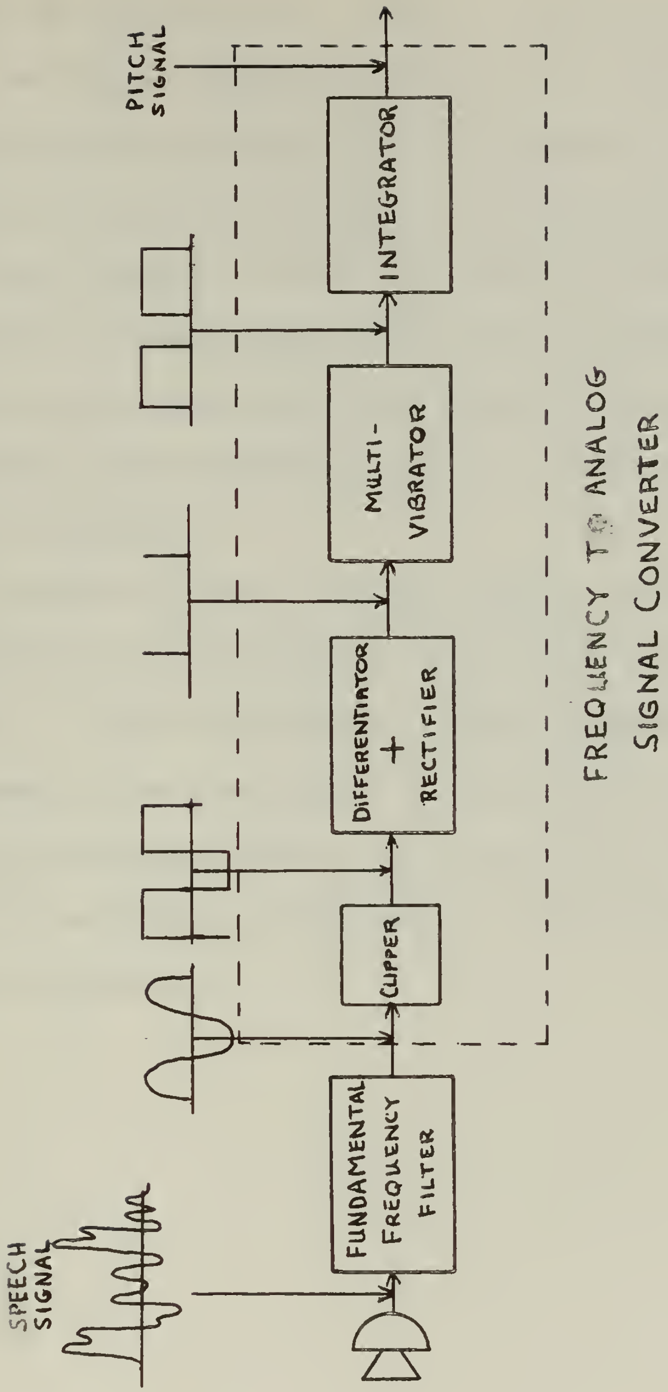


Figure 3-2. Block diagram of pitch extractor channel.

[illegible]

[illegible]

fundamental is transmitted. However, when the sound is unvoiced, the pitch extraction channel is simply deenergized and this fact is transmitted to the vocoder receiver. As stated earlier, sounds which are a combination of voiced-unvoiced, most vocoders treat the sound as voiced. This generalization includes many sounds and therefore, we must find better ways to transmit more specific information if we are to achieve higher intelligibility scores on unvoiced or semi-unvoiced sounds.

In a recent test /6/, intelligibility scores between 60 and 85 percent were obtained for the fixed channel vocoder (Figure 3-3). This corresponds to 90% sentence intelligibility and this seems to be acceptable for military use (Figure 3-4). However the compression ratio is about 10:1, and as stated earlier, with suitable multiplexing, little reduction in bandwidth is realized.

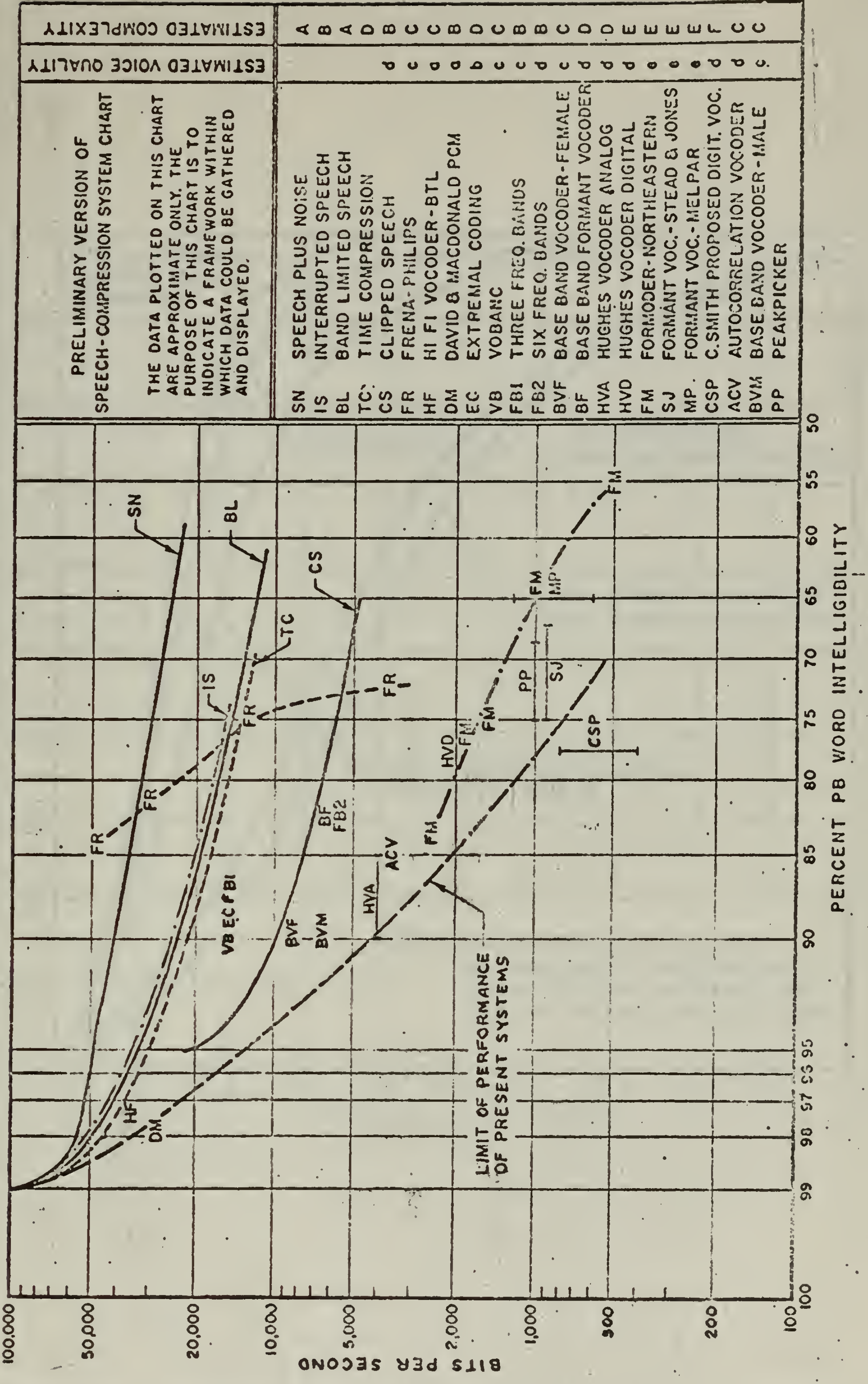


Figure 3-3. Relative performance of present-day speech compression systems. (Est. V. Qual. Range; a-f: Est. Complex. Range; A-F)

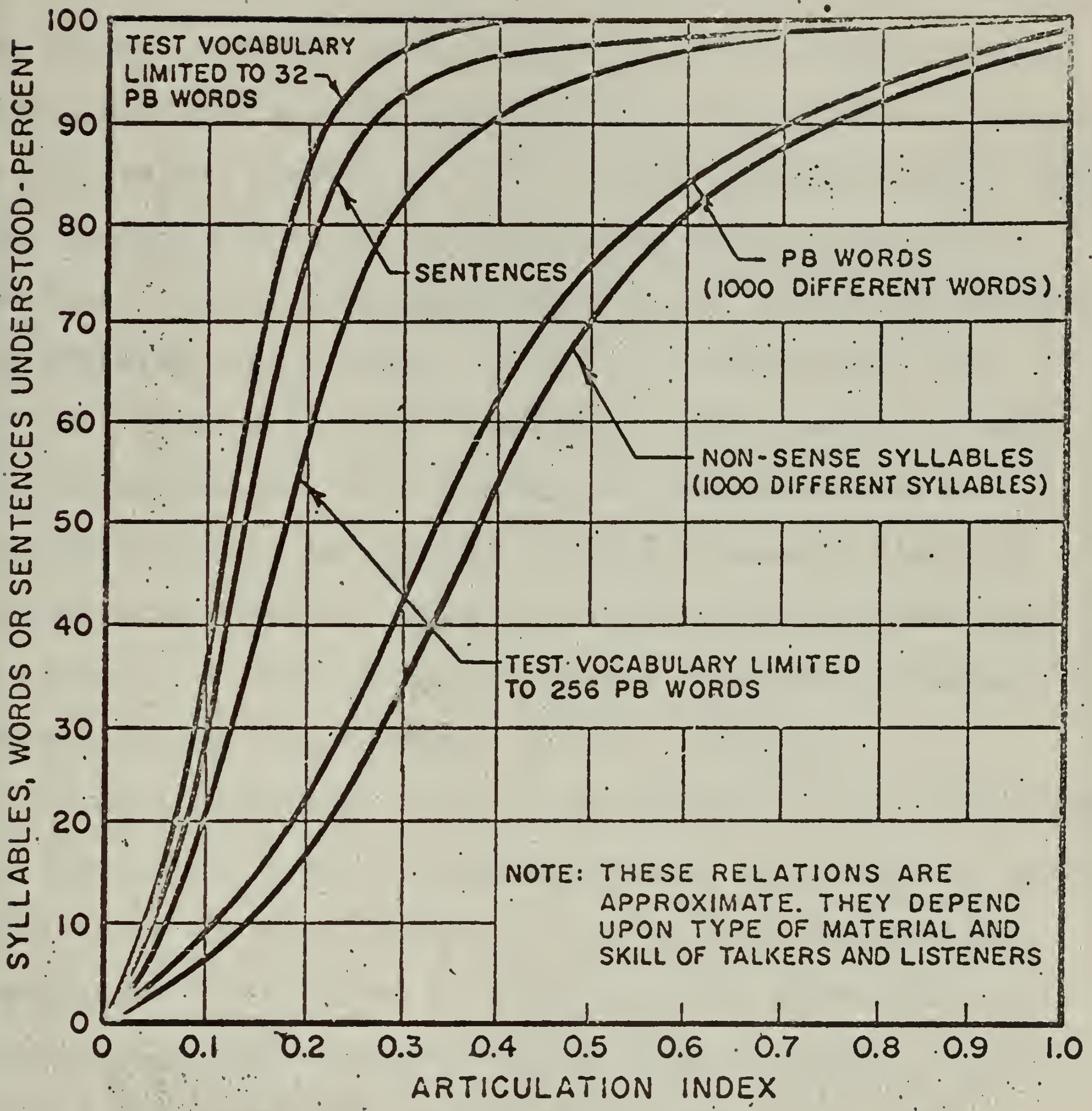
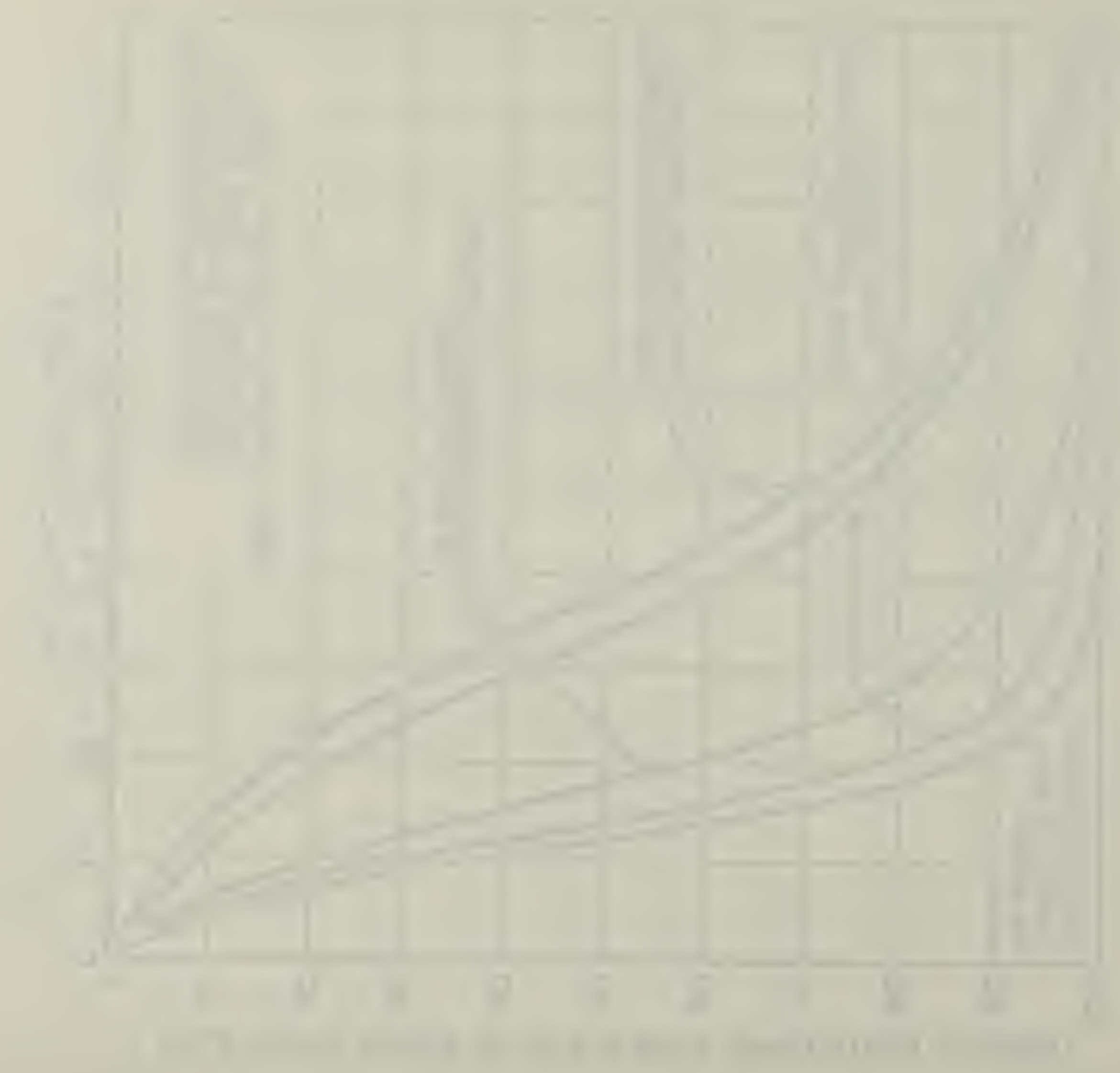


Figure 3-4. Conversion chart for predicting test scores given results of one test.



4. Correlation Vocoders

The ultimate requirement of this vocoder technique, as in the channel vocoder, is to correctly reproduce the spectral envelope of any signal input between zero and four kilocycles. As pointed out previously, the spectrum channel vocoder commits some serious errors in its reproduction of the spectral envelope, thereby reducing the quality of speech obtainable. Specifically, these errors are due to the fact that the filters, because of their bandwidths, individually average several harmonics. This problem is compounded when one considers that it is possible for the fundamental and its harmonics to slide to adjacent filters during an utterance. As a result, the filter bank analysis of speech is much more erratic than a true harmonic (Fourier) analysis.

Studies conducted by M. R. Schroeder /13/ have indicated that these errors may be greatly reduced by performing speech analysis in the time domain, i. e., by using correlation techniques. This theory is based upon the Wiener-Khinchine theorem /14/ which states that the autocorrelation function and the power spectral density of a given signal are a Fourier transform pair. Since the theorem is based on integrating in time from minus infinity to plus infinity, the relationship must be extended to short-time analysis. This was done by R. M. Fano /15/ and his theory was applied by R. Biddulph /16/. M. R. Schroeder has simulated time-domain vocoders based on this correlation principle /17, 18/.

4.1 Autocorrelation Vocoder

The autocorrelation function, $\phi(T)$, of a signal, $s(t)$, is defined as follows:

$$\phi(T) = \overline{s(t) * s(t-T)},$$

where the bar indicates a time average, which, for speech is taken over a 30 ms period. This is an even function and is bandlimited to the same frequency range as the signal itself. The transform of the autocorrelation function gives a spectrum which is the absolute square of the signal spectrum. Therefore, it contains the same information.

A complete autocorrelation vocoder is shown in Figure 4-1. In the analyzer, the short-time autocorrelation function of the signal is derived for a number of discrete delays, T_o, T_1,-----, T_n. Since the function is bandlimited to the same frequency band as the speech signal, the function is completely defined by discrete delays of $1/2f_c$ where f_c is the cut-off frequency of the speech. Therefore, if we consider f_c to be 3kc, a $\Delta T = .167$ ms will suffice. The maximum delay for which the short-time autocorrelation needs to be specified is of the order of 3 ms, or about a pitch period. Therefore, a total of 18 delay channels are required, each 20 cps wide, for a total bandwidth of 360 cps. A compression ratio of about 9:1 is obtained but pitch information must be included in addition. Thus, we have obtained speech compression of the same ratio as the channel vocoder. We might have expected this due to the fact that both methods obtain

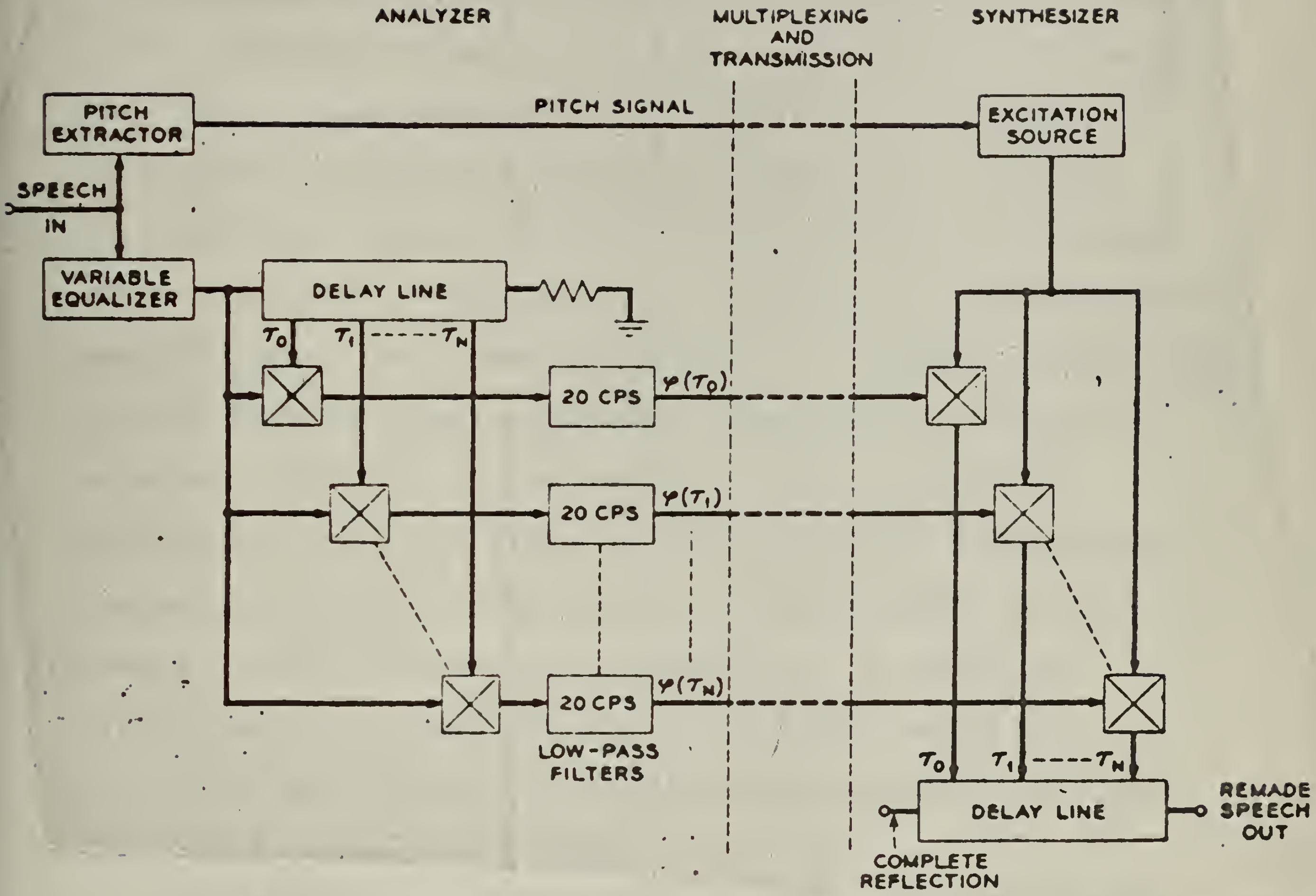


Figure 4-1. Block diagram of an autocorrelation vocoder.

compression by discarding phase information.

At the synthesizer, a symmetrical replica of the auto-correlation function is generated for every pitch period. This is achieved by a simple reciprocal scan giving a synthesized signal whose amplitude spectrum is the square of the original speech spectrum. This spectrum squaring must be compensated for, and so, the addition of an equalizer is indicated. A time-varying equalizer is shown in Figure 4-2. As shown, the equalizer consists of three filters (for the separation of the formants), rectifiers, low-pass filters, square rooters and dividers. The formant amplitudes are reduced to the square root of their original amplitudes, compensating for the squaring introduced in the analysis. This equalizing operation is very difficult to realize without very complex equipment and this is the reason that this method of voice coding has not received wide acceptance. In addition, distortion due to the chopping of the pitch period is introduced and is quite noticeable in the synthesized speech.

4.2 Cross-correlation Vocoder

The spectrum squaring problem may be avoided by cross-correlating the original speech signal with a speech-derived signal having a flat spectral envelope over the band desired (0-3kc). The synthesizer of a cross-correlation vocoder is identical to that of the auto-correlation vocoder without the equalizer, provided that the cross-correlation function is reasonably symmetric so that reciprocating scanning may be

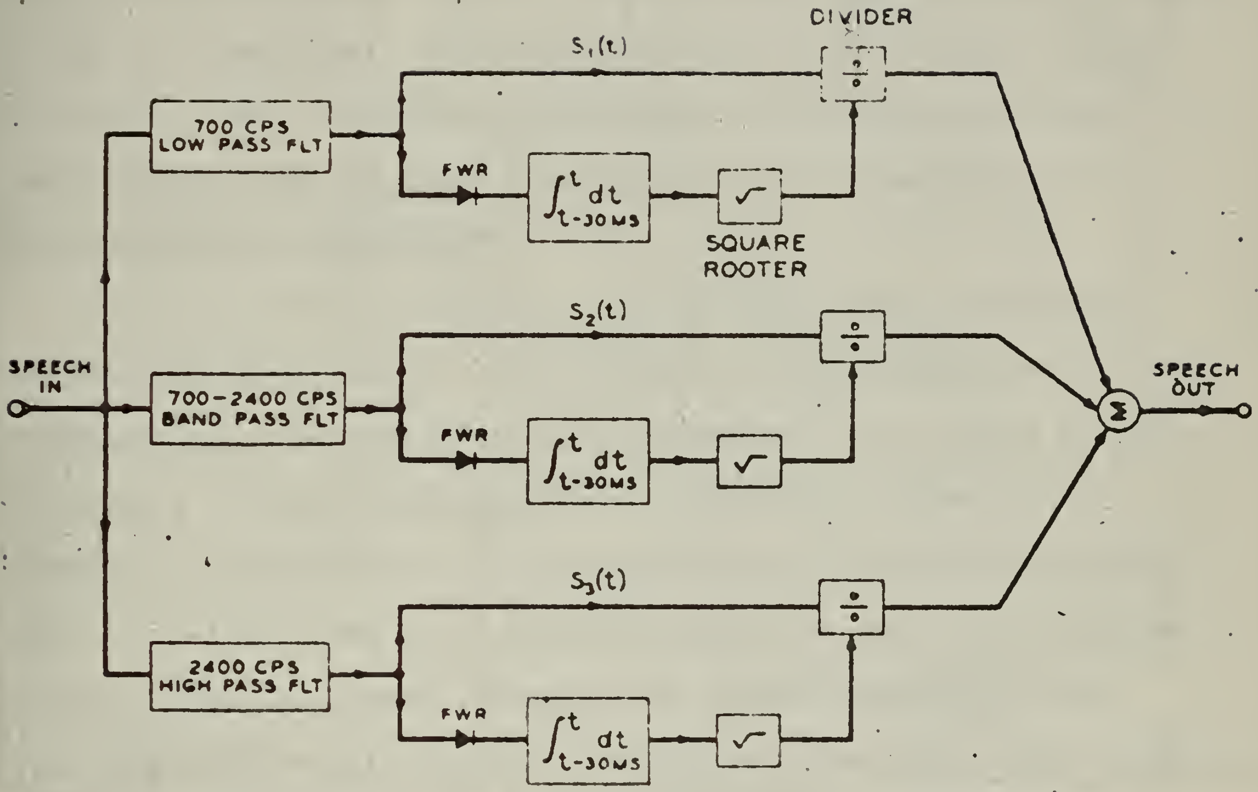


Figure 4-2. Block diagram of a variable equalizer.

employed. The block diagram of a cross-correlation analyzer is shown in Figure 4-3.

At this time, cross-correlation vocoders are somewhat inferior to auto-correlation vocoders since the cross-correlation function cannot be made truly symmetric or be made to have a completely flat spectrum over the frequency range involved. The outstanding advantage of cross-correlation analysis is that no analog multipliers are required as in auto-correlation analysis.

In conclusion, it may be stated that while correlation techniques have given a new approach to the problem of speech compression, the operations indicated call for fairly complex equipment of very good quality. In addition, there is no reason to suspect that intelligibility or articulation scores will be better than in the fixed channel vocoder (See Figure 3-3). It is for these reasons that these techniques have received little support from the military and the only complete working vocoder based on this principle was simulated on a computer /13/.

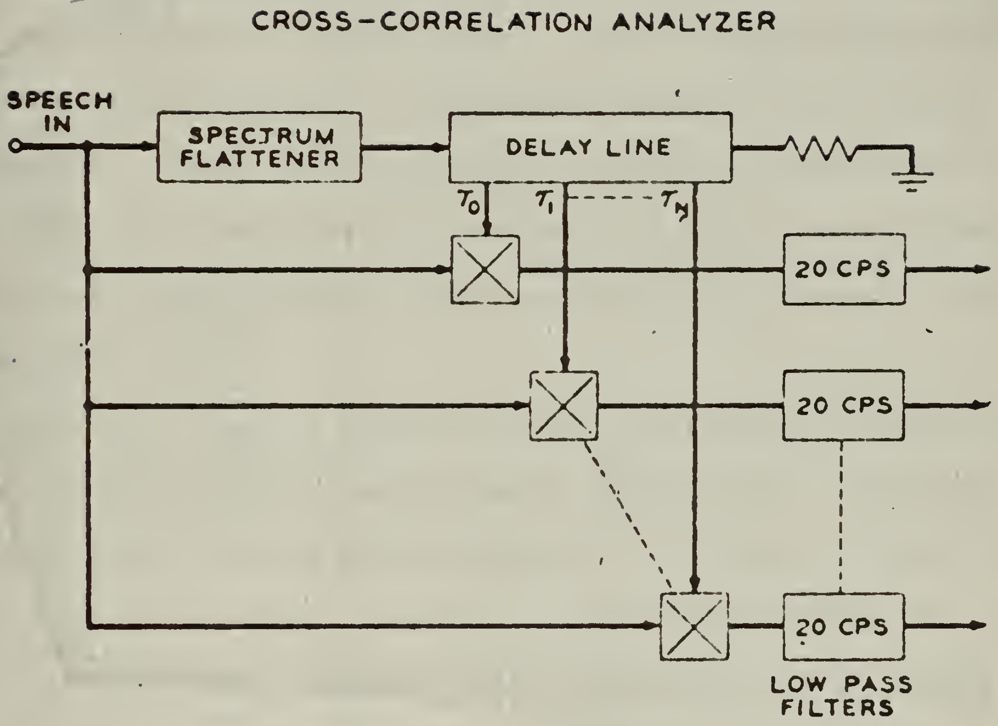


Figure 4-3. Block diagram of a cross-correlation analyzer.

5. Formant Tracking Vocoder

It has been known for some time that a unilateral relationship exists between the configuration of the vocal tract and the formant frequencies, and a bilateral relationship exists between the perception of certain speech sounds and the formant frequencies /19/. Therefore, the formant frequencies appear to constitute a set of nearly independent, slowly varying parameters that convey much of the speech information. If it were possible to devise a system in which only the slowly varying parameters were transmitted, an extremely narrow band speech communication channel could be designed.

Typically, such a channel would incorporate an analyzer at the transmitting end which would extract the important characteristics of the speech signal and transmit them to the synthesizer which would reconstruct the speech from the control data. In this case, however, the synthesizer is a lumped constant network that exhibits transmission properties similar to the transmission properties of the vocal tract. That is, the analyzer accepts continuous speech as an input and automatically supplies control data which alter the excitation and transfer properties of the "terminal-analog" network so that its transduced output resembles the original speech /20/.

A typical analyzer is shown in Figure 5-1. In this case, the parameters extracted are:

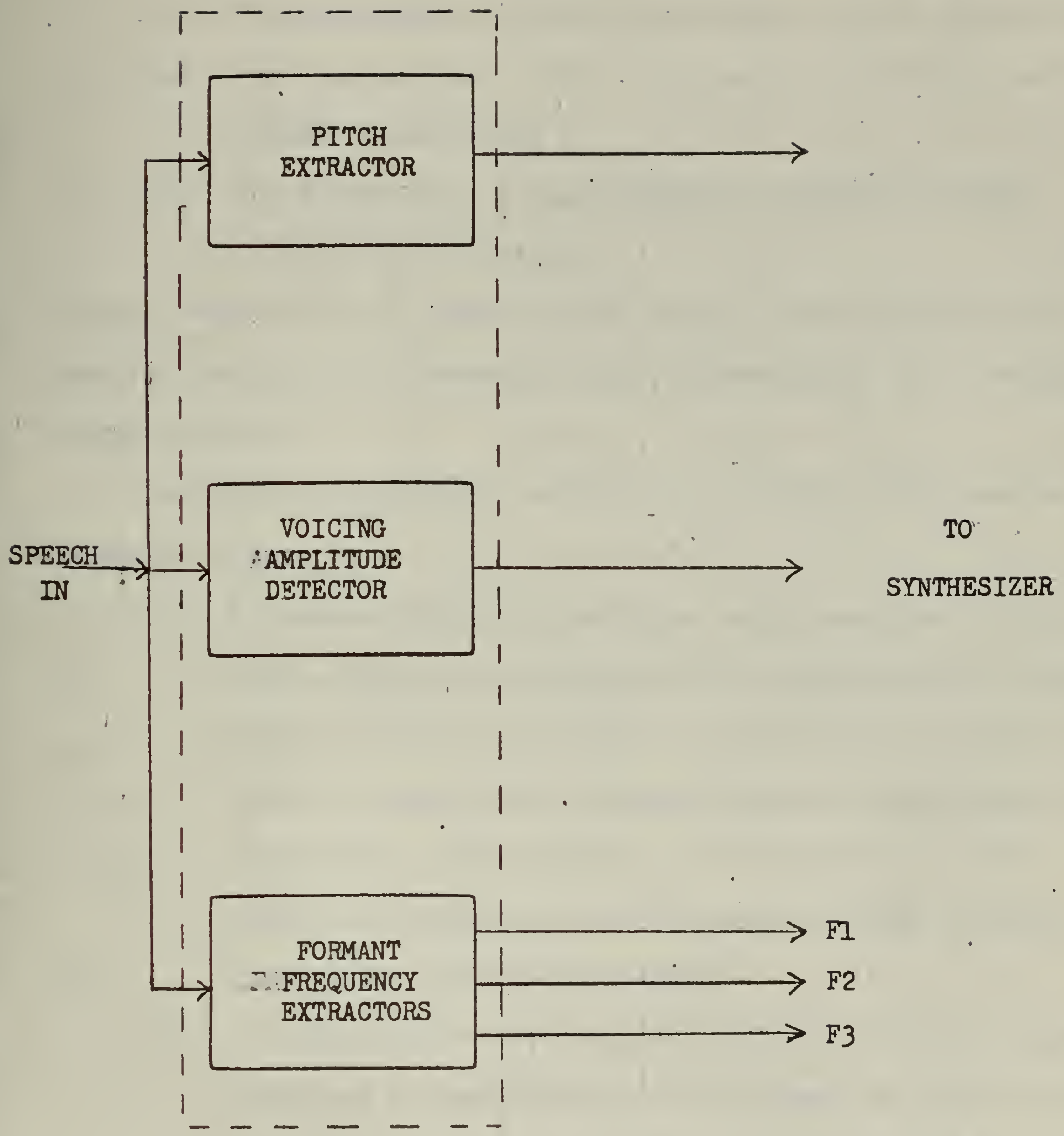


Figure 5-1. Block diagram of a formant tracking analyzer.

(1) The first three formant frequencies;

(2) The amplitude of voicing;

(3) The amplitude of friction (noise, as in fricatives);

(4) The fundamental vocal frequency (in vowels and voiced consonants);

(5) The frequency of the spectral maximum of the fricative excitation.

A total bandwidth of about sixty cycles per second is all that is required to transmit this information, or a reduction of about 50:1!

The basic components required to extract these seven parameters are:

(1) Formant-tracking apparatus which accepts natural input speech and yields four slowly-varying voltages, three of which represent, as functions of time, the values of the first three formant frequencies, and the fourth representing the frequency of the spectral maximum in the frequency range usually containing fricative energy;

(2) A voicing-friction (noise) detector which yields voltages approximately proportional to the amplitude of voicing and of fricative excitation of the vocal tract during speech production.

(3) A pitch extractor circuit which produces a voltage proportional to the fundamental frequency of voiced sounds. A typical vocal-frequency

indicating circuit is shown in Figure 5-2. For a detailed description of these circuits, see /37/ and /21/. All voltages extracted are low-pass filtered prior to transmission.

As explained above, the synthesizer utilized in this system is an analog of the vocal tract controlled by the transmitted parameters. The vowel-producing portion of the synthesizer is composed of four cascaded, uncoupled simple resonant circuits excited from a source of repetitive pulses as shown in Figure 5-3. Balanced reactance modulators which are electrically variable are utilized as the capacitive elements of the first three resonators. Consonant sounds are synthesized by supplementing the vowel-producing circuitry with a noise source whose amplitude and spectral properties are controlled by the analyzer. An example of this is shown in Figure 5-4. A detailed description of the synthesizer circuitry is given in /22, 23/.

With respect to the factors of prime importance to be exhibited by a speech compression system for broad military use, the amount of compression obtained with the formant tracker is adequate. However, intelligibility scores obtained with the standard Harvard PB word lists are not acceptable, averaging about 33%. In addition, the only formant tracker in production, built by Melpar, did more poorly on all other tests than did six other types of vocoders /6/ (See Figure 3-3). These results have rendered the formant tracker, in

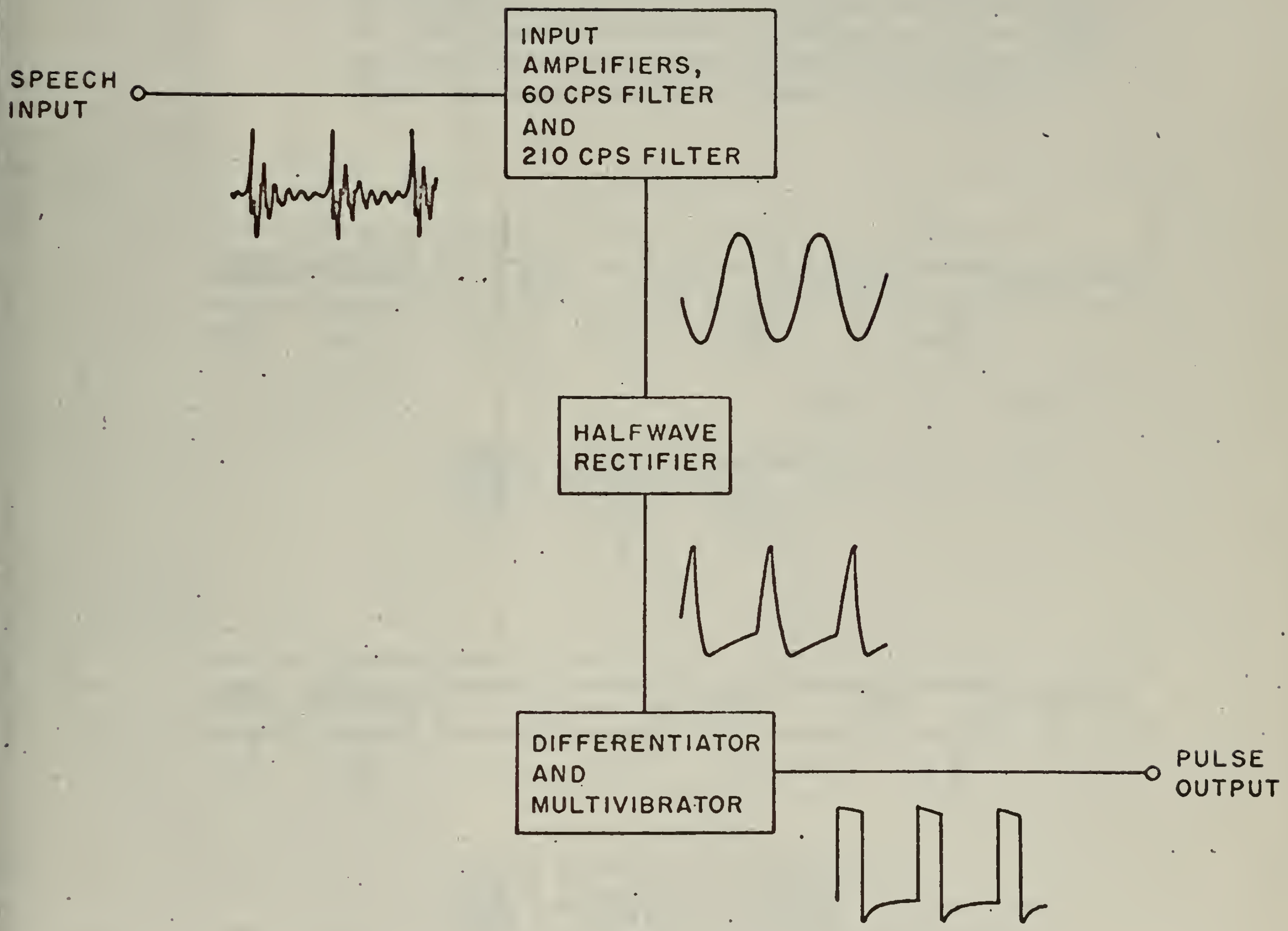


Figure 5-2. Block diagram of a vocal frequency indicator.

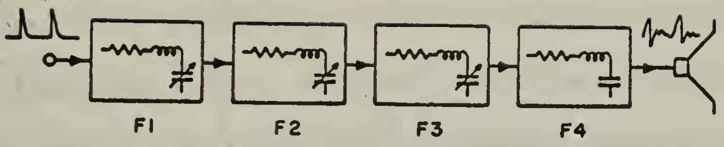


Figure 5-3. Block diagram of a terminal-analog vowel synthesizer employing cascaded resonators. An impulsive source of excitation analogous to the glottal source is shown at the left, and the resonators are labeled to indicate their association with speech formants.

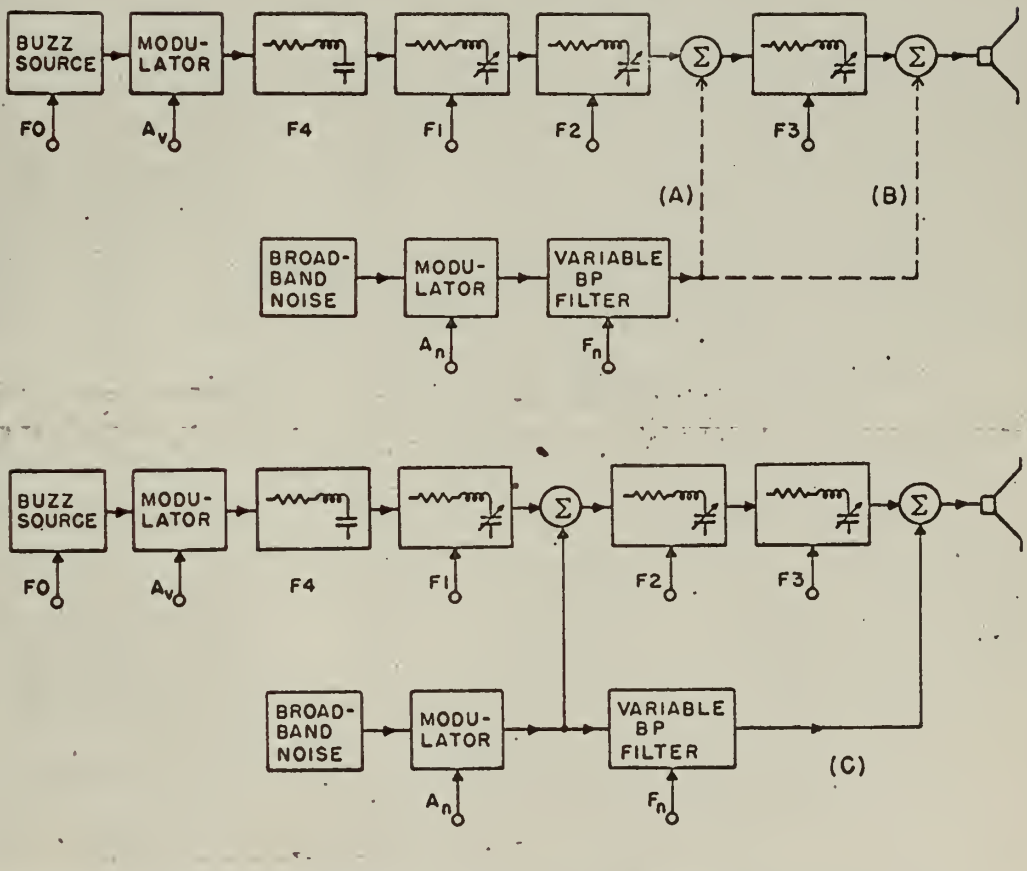


Figure 5-4. Block diagram showing three modifications of the basic vowel synthesizer. These modifications represent empirical attempts to produce consonant sounds.

its present form, useless to the military. However, much work has been done by Flanagan of Bell Telephone Laboratories and others to try to improve system response to consonant sounds. Their work has resulted in several modified systems, the most important being the resonance vocoder described in the next section.

6. Hybrid Vocoders

There are so many variations of the three types of vocoders just discussed that it would be too voluminous to discuss them all in this paper. However, a few of these hybrids are receiving a considerable amount of attention, and these will be briefly discussed.

The most unimaginative speech compression technique is to transmit intact that portion of the speech wave which gives the most trouble to the basic vocoder system you are studying. For example, in the formant tracker proposed by Flanagan, one of the most flagrant sources of error is the formant tracking circuitry. This circuitry operates very badly if the rate of speech is increased or if the formants slide, as happens when the speech contains an acoustic "glide"[1]. Flanagan proposed that these problems would be eliminated if the band from 300 to 710 cycles per second (commonly called the "baseband") were transmitted intact, i. e., by analog methods, and the remaining spectrum represented by three resonances requiring a total of six control signals. In the synthesizer, this baseband is used as the exciter source and is modulated by the control signals. Bell Labs found that much better speech quality could be obtained if the baseband were put through a non-linear device prior to modulation as higher harmonics of the fundamental frequencies were generated in this manner giving a broader, more realistic spectrum. The

[1] See Appendix A for definition.

compression ratio is lowered to about 5:1 by this method, but sentence intelligibility is up to 90 per cent (Figure 3-3). This particular equipment is called, "Voice-Excited Resonance Vocoder" /7/. The compression ratio obtained is unacceptable for wide military acceptance.

Another variation is similar in technique to the fixed channel vocoder and is called the "Peakpicker". The idea incorporated in this device is to conserve bandwidth by restricting the transmission to those channels that are carrying the most information at each instant. The Peakpicker selects the three (or four or five) largest spectrum peaks that it can find at a particular instant and transmits to the synthesizer only this information. A variation of this idea is called "Vosdic". Each channel has a certain set threshold which when exceeded, is indicative of information in that band. A simple scanner is employed and ones or zeros are transmitted to the synthesizer corresponding to the instantaneous energy levels in each channel /24/.

7. Statement of the Speech Compression Problem

The significant speech compression devices have been described in considerable detail. To reiterate, the central problems of the analysis-synthesis scheme of speech compression are:

(1) Determine the most suitable form in which to code the speech signal, such that the channel capacity required for transmission of the coded signal is as low as possible;

(2) Develop instrumentation to extract the data in (1) with minimum error, and to receive the coded data and synthesize acceptable speech.

Problem (1) is simply the problem of speech recognition. It is the author's belief that a greater effort should be extended in solving this problem, especially when the greatest number of such devices will be utilized in the tactical military environment where the amount of compression and intelligibility are the factors of utmost importance.

In retrospect, it should be remembered that communications systems designed to transmit information generally fall into two categories:

(1) Those for which the input is a continuum, and the output, derived through some transformation, is the best possible imitation of the input. Even though the range of values of the input is limited, an attempt must be made to produce a unique output

for every value of the input in that range. Therefore, the number of values possible at the output approaches infinity as the quality of the transmission increases. Radios, phonographs, and tape recorders are examples of this type of system.

(2) Those for which the input can be expressed in terms of a fixed set of discrete values. The input is considered encoded into members of this set, while the output may only be a repetition of the input rather than a transformation. Again, there is an output symbol for every value of the input. However, if the input is bounded, only a finite number of output values will serve to distinguish among all possible inputs. Pulse code modulation and other digital techniques fall into this category.

It is evident that no matter how small the imperfections in the individual links of type 1, a sufficient number in cascade will produce a system in which there is no measureable correlation between output and input. In contrast, however, if the imperfections in the links of type 2 are only small enough not to cause a given input to produce more than one of the discrete outputs, compound systems will perform perfectly regardless of the number of links.

If we are to have repeatability, the set in terms of which all messages are to be represented must be denumerable.

One of the most important characteristics of the speech process is the preservation of repeatability as a message is communicated from one speaker to another. It is important to note, however, that in no sense is the speech signal representing a message reproduced exactly as it is passed from speaker to speaker. Therefore, we may expect to find a code or discrete and finite set of units common to all speakers of a language which will suffice to represent any speech event recognizable as part of a language /25/.

Following a review of speech recognition (encoding) techniques, the writer has reached the following conclusions:

(1) Most techniques operate on the electrical sound signal as if it were a continuous electrical signal whose frequency components lie in the audio band;

(2) Speech researchers have been reluctant to consider the phonemic level of the language as the proper recognition level, for they have felt that it was not possible to extract phonemes from continuous speech with sufficient accuracy. They have preferred to work at the word or syllable level as the statistical variations in the basic elements, due to different speakers or context, are less;

(3) Ten years ago, Jakobson, Fant, and Halle /1/ pointed out that there are certain distinctive linguistic features which should have electrically measureable counterparts, which, if measured,

could uniquely define phonemically, what was said;

(4) Heretofore, speech engineers have been content with the rather elementary coding scheme of detecting only a small number (perhaps one or two) of the fundamental linguistic features and employing other coding procedures for the remaining components of the signal. For example, the presence or absence of voicing is essential to most existing systems, but the remaining information is obtained in a myriad of ways.

Conclusions (1), (2), and (4) have limited the amount of compression attainable with existing systems. Particularly in (2), the space of input events is greatly increased if any other than the phonemic level is considered, thereby greatly decreasing the maximum attainable compression ratio. The desire for naturalness and speaker recognition capability has also limited the amount of compression. Based on the recent work of Hughes and others /25-28/, it is felt that phoneme recognition is definitely possible and that the recognition technique utilizing distinctive features, is more promising than any other which has been studied. In summary then, a phonemic code contains a "finite set of units common to all speakers of a language which will suffice to represent any speech event recognizable as part of a language".

Having established the desirability of phonemic coding, let us make some elementary observations concerning speech. Continuous speech may be regarded as a sequence of discrete events, or phonemes, selected from an ensemble of, say, 35 phonemes (in English; Appendix B) occuring at the rate of about ten per second. It has been established that these events are neither equiprobable nor independent, therefore, the information rate is about 50 bits per second. Such a sequence of phonemes, if properly synthesized with appropriate transitions, could be quite intelligible, although there would be a loss in naturalness. One concludes, therefore, that a speech-compression system based on the coding of phonemes would require a channel capacity of about one five-hundredths of that required for a conventional speech channel. Thus, the realization of circuitry based on the "method of distinctive features" will provide a speech communication system whose information rate is commensurate with the information rate of speech.

8. The Method of Distinctive Features

Phonemes may be distinguished from one another only by their complex differences and not by unitary properties. These complex differences, i. e. distinctive features, however, do have individual uniqueness and may always be distinguished from each other by a single binary decision.

The "method of distinctive features" states that any phonemic utterance may be identified by the resolution of a series of opposition pairs. Within a given language, each pair of oppositions has a specific property which differentiates it from all others. The distinctive features are the ultimate entities of a language since no one of them can be broken down into smaller linguistic units. A simultaneous group of distinctive features forms a phoneme; that is, for a given language, a phoneme is uniquely specified when the binary decisions are made which correspond to each of the distinctive features.

Distinctive features may be divided into two classes; 1) prosodic and, 2) inherent. Prosodic features are displayed only by those phonemes which participate in the intonation pattern and may be defined only with reference to a time series. On the other hand, inherent features are displayed by all phonemes and are definable without reference to a sequence. These are the features that we wish to define and measure.

The inherent distinctive features which have been discovered thus far, amount to 12 oppositions, from which each language makes its own selection. For English, ten oppositions are evident and they are defined in an acoustical and genetic sense in Figure 8-1.

The inherent distinctive features necessary to specify the 35 phonemes required of English are listed in Figure 8-2a along with the phonemes to which they apply. Blank spaces in Figure 8-2a indicate that the feature is not required to distinguish a given phoneme uniquely from the others. Figure 8-2b is a "tree" diagram which indicates the manner in which the successive binary decisions of Figure 8-2a uniquely specify each phoneme /28/.

The description of the distinctive features presented in Figure 8-1 is a summary and is intended only to indicate the way in which a phonetician describes the co-ordinates of the physical space defined by the distinctive features. The matter of how to describe this space in engineering terms is discussed in the next section. As of this time, experimental work in this area has shown that measurements based on these definitions have, for several of the features, yielded good results /29/. However, it must be emphasized that these definitions are useful only as guides to the required measurements, and it is the objective of the proposed research outlined herein to determine these measurements as related to their acoustical correlates. Additionally, it

Correlates of Distinctive Features

Distinctive Features	Acoustical Correlates	Genetic Correlates
1) **Vocalic/Nonvocalic***	Presence versus absence of a sharply defined formant structure.	Primary or only excitation at the glottis together with a free passage through the vocal tract.
2) **Consonantal/Nonconsonantal**	Low versus high total energy.	Presence versus absence of an obstruction in the vocal tract.
3) **Compact/Noncompact**	A maximally high first formant.	The vocal tract assumes a position in which the cross-sectional area increases (or at least does not decrease) forward from the glottis to the lips.
4) **Diffuse/Nondiffuse**	A maximally low first formant.	The vocal tract assumes a shape similar to that of a Helmholtz resonator, that is, there is a constriction forward in the oral cavity.
5) **Grave/Acute**	Concentration of energy in the lower (versus upper) frequencies of the spectrum.	Peripheral versus medial: peripheral phonemes (velar and labial) have more ample and less compartmented resonator than the corresponding medial phonemes (palatal and dental).
6) **Continuant/Interrupted**	Silence (at least in the frequency range above the vocal cord vibration) followed and/or preceded by a spread of energy over a wide frequency region (either as a burst or as a rapid transition of vowel formants) versus absence of abrupt transition between sound and "silence."	Rapid turning on or off of source either through a rapid closure and/or opening of the vocal tract that distinguishes plosives from constrictives or through one or more taps that differentiate the discontinuous liquids like a flap or trill /r/ from continuant liquids like the lateral /l/.
7) **Flat/Plain**	Flat phonemes are opposed to the corresponding plain ones by a downward shift or weakening of some of their upper frequency components.	The flat (narrowed slit) phonemes, in contradistinction to the plain (wider slit) phonemes, are produced with a decreased back or front orifice of the mouth resonator, and a concomitant velarization expanding the oral cavity.
8) **Nasal/Oral**	Spreading the available energy over wider (versus narrower) frequency regions by a reduction in the intensity of certain (primarily the first) formants and introduction of additional formants.	Oral resonator supplemented by the nasal cavity versus the exclusion of the nasal resonator.
9) **Tense/Lax**	More (versus less) sharply defined resonance regions in the spectrum, accompanied by an increase (versus decrease) of the total amount of energy and its spread in time.	Greater (versus smaller) deformation of the vocal tract away from its rest position.
10) **Strident/Mellow**	Higher intensity noise versus lower intensity noise.	Rough-edged versus smooth-edged; supplementary obstruction creating edge effects at the point of articulation distinguishes the production of the rough-edged (strident) phonemes from the less complex impediment in their smooth-edged (mellow) counterparts.

* This feature has been replaced by the feature Sonorant/Nonsonorant as explained in Section 9.1 of the text.

Figure 8-1. Correlates of distinctive features.

	PHONEMES																																		
DISTINCTIVE FEATURES	i	ɪ	u	ʊ	ɛ	e	æ	o	ɔ	a	ʌ	r	l	w	j	m	n	ŋ	f	s	ʃ	θ	ð	v	z	ʒ	ʤ	ʧ	p	t	k	b	d	g	h
1) SONORANT / NONSONORANT	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
2) CONSONANT / NONCONSONANT	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
3) COMPACT / NONCOMPACT						−	+	−	+	+						−	−	+	−	−	+	−	−	−	−	+	+	+	−	−	+	−	−	+	
4) DIFFUSE / NONDIFFUSE	+	+	+	+	−	−	−	−	−	−	−																								
5) GRAVE / ACUTE	−	−	+	+	−	−	−	+	+	+	+	−	+	+	−	+	−		+	−		+	+	+	−				+	−		+	−		
6) CONTINUANT / INTERRUPTED																			+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	
7) FLAT / PLAIN									+	−		+	−	+	−																				
8) NASAL / ORAL												−	−	−	−	+	+	+																	
9) TENSE / LAX	+	−	+	−	−	+	+	+	+	+	−								+	+	+	+	−	−	−	−	−	+	+	+	+	−	−	−	
10) STRIDENT / MELLOW																			+			−	−	+			+	+			−			−	

Figure 8-2a. Phonemes of English and their distinctive feature composition.

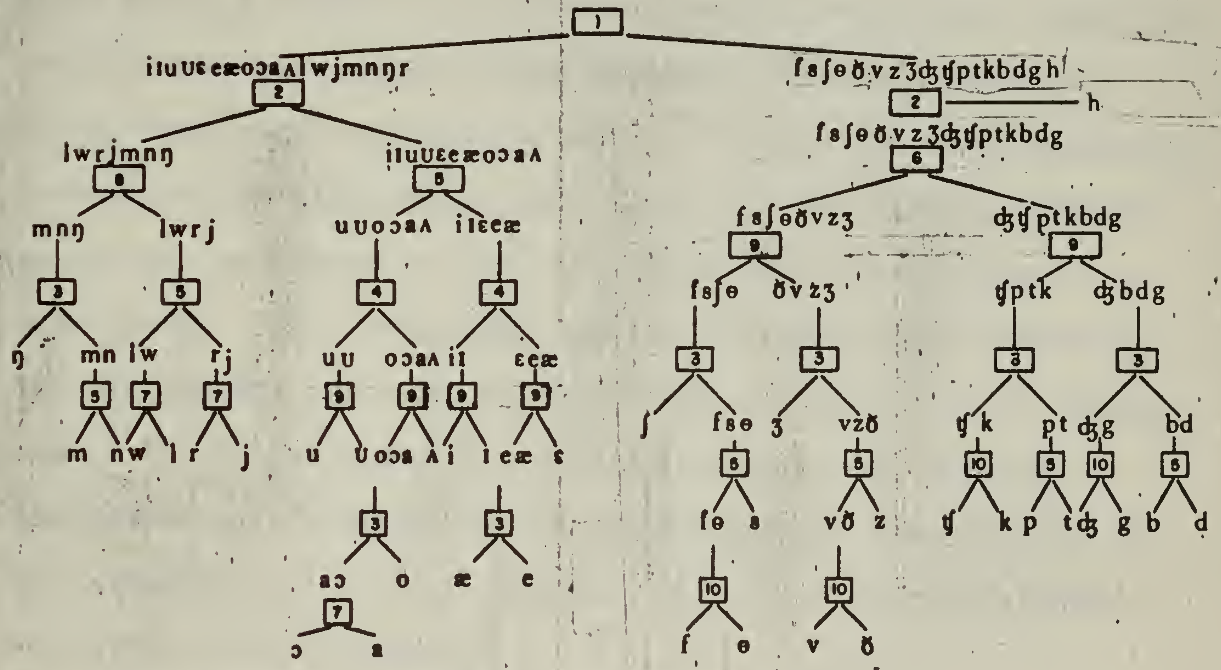


Figure 8-2b. Tree diagram for the identification of phonemes.

may be that an empirically derived set of measurements may yield information equivalent to the determination of a particular feature, and therefore, would be of greater value.

The complete physical determination of acoustical correlates would resolve the speech recognition problem, but it is interesting to note that even a partial solution (such as occurs when one or more of the distinctive feature measurements cannot be made) is significant in that each distinctive feature selection separates the ensemble of phonemes into predictable subsets. The importance of this fact is twofold: 1) a large limited vocabulary could be constructed using only those phonemes which may be determined with high probability and, 2) errors in intelligibility scoring which might occur due to one or two inexact measurements may be traced to one or a small set of phonemic utterances. This reasoning may be substantiated intuitively by considering the effect of deleting one-third of a printed text. The intelligibility of that text is not reduced by one-third and furthermore, the reduction in intelligibility is lessened in this case because we have some statistical knowledge of the mistakes.

The distinctive features may be divided into two broad classes depending on the amount of accumulated knowledge concerning them. The first class contains those features whose acoustical correlates have been reasonably well determined. The second class contains the less well specified

features requiring future research. The section following is a discussion of the features of the first class.

9. Specified Features

9.1 Sonorant/Nonsonorant

This feature was invented by G. W. Hughes /25/ in lieu of the feature vocalic/nonvocalic as originally proposed by Jakobson, Fant, and Halle /1/. The acoustical correlate of the feature vocalic/nonvocalic is the presence versus the absence of a sharply defined formant structure. The purpose of the feature was to separate vowels and liquids from the nasals, affricates, fricatives and stops. While it is quite easy to determine the presence or absence of a sharply defined formant by visual inspection of a sonagraph of the phoneme, it is more difficult to make this determination by examining the discrete amplitude spectrum of the time waveform of the phoneme in a computer. Due to this difficulty, Hughes used the feature sonorant/nonsonorant to separate vowels, liquids, and nasals from the affricates, fricatives, and stops. This technique shows considerable promise, since the separation can be made by straightforward spectral measurements, and the nasals are then separated from the vowels and liquids on the basis of the nasal/oral distinction. Hughes concluded from his experiments that this feature may be determined by comparing the energy measured below 1 kc (excluding the energy contributed by the pitch fundamental), i. e. "voicing", with that measured above 2.5 kc. Sonorant sounds exhibit greater energy in the lower band.

9.2 Consonantal/Nonconsonantal

Consonantal sounds are generated by obstructing the vocal tract in various ways thus causing a drop in the radiated acoustic energy during the transition from nonconsonantal to consonantal sounds. Note that there may be a distinct formant structure, as in the case of liquids, or, the obstruction may cause turbulent noise, as in the case of fricatives, or both. In any event, the drop in radiated energy is the clue. The feature consonantal/nonconsonantal will therefore be determined by comparing the energy measured above 300 cps (to exclude voicing) at a given instant of time with the energy measured in the same manner at an earlier time (about 50 ms) /30/.

9.3 Compact/Noncompact

In order to measure the acoustical correlate of the feature compact/noncompact, phonemes must first be separated into three classes: 1) sonorants, 2) stops, and 3) fricatives and affricates.

(1) For sonorant sounds, compactness is evidenced by a maximally high first formant. In particular, the first formant of compact sonorant phonemes is generally greater than 600 cps.

(2) In describing the measurements to be performed in determining the feature compact/noncompact for stops, it is important that the grave/acute decision be made previously. This decision

divides the eight English stops (i. e. /t/, /p/, /d/, /b/, and the front and back variants of /k/ and /g/) into two groups. One group consists of the grave stops /d/ and /b/ and back variants of /k/ and /g/, while the other group is made up of the acute stops, /t/ and /d/, and the front variants of /k/ and /g/. The first group is finally divided into compact/noncompact by comparing the outputs of the passbands which are 1kc to 2kc and 300 cps to 7kc. This measurement as determined by Halle, Hughes and Radley /31/ indicates that the back variants of /k/ and /g/ (which are compact) have a higher average energy in the 1kc to 2kc band relative to the 300 cps to 7kc band, than do /p/ and /b/. Similarly, the front variants of /k/ and /g/ (which are compact) have a higher average energy in the 2kc to 4kc band relative to the 300 cps to 7kc band, than do /t/ and /d/.

(3) For fricatives and affricates, compactness is indicated by the presence of sharp resonances in the band 1.7kc to 3.4kc. Therefore, a comparison must be made between the acoustic energy in the 1.7kc to 3.4kc band and the acoustic energy in the 700 cps to 1.4kc band. The threshold involved must be determined after experimental data has been collected on the particular equipment being used.

9.4 Diffuse/Nondiffuse

Diffuse vowels are characterized by a concentration of energy in the lower portion of the spectrum. Peterson and Barney /32/ found that the first formant of diffuse vowels rarely exceeds 500 cps; therefore, this is the threshold.

9.5 Grave/Acute

Again, it is necessary to describe the measurements required for the determination of this feature for three classes of phonemes: 1) sonorants, 2) fricatives and affricates, and 3) stops.

The experiments of Peterson and Barney /32/ provide the basis for the grave/acute measurement for sonorant sounds. Jakobson, Fant, and Halle /1/ stated that the acoustical correlate of this feature is evidenced by the relative concentrations of energy in the lower versus the higher frequencies of the spectrum. Table 9-1 devised from the data of Peterson and Barney is shown below:

TABLE 9-1

$F_2 - F_1$ for English Vowels

Grave					Acute			
u	U	ɔ	a	ʌ	i	I	ɛ	æ
500	650	200	300	600	2200	2000	1600	1100

The procedure is to subtract the first formant frequency, F1, from the second formant frequency, F2, and if this

difference is less than 850 cps, then the sonorant in question is grave.

For fricatives and affricates, different measurements must be performed. In particular, a comparison must be made between the energy in the 4kc to 7kc band and that in the 700 cps to 7kc band. A threshold ratio, depending on the equipment involved, will have to be set, below which, the sound would be classified as grave.

To determine whether stops are grave or acute, Halle, Hughes, and Radley /31/ have shown that acute stops have energy concentrations in the high frequencies, whereas grave stops, have weak high frequency components. While this is very similar to the basis for the fricative and affricate grave/acute decision, experimentation has shown that a comparison of the energy in the 2.5 kc to 7 kc band relative to that in the 700 cps to 7 kc, with the appropriate threshold, gives better results for stops than the ratio indicated for fricatives and affricates.

9.6 Continuant/Interrupted

Interrupted sounds are caused by closure or near closure of the vocal tract at some point. The acoustical correlate of this is, therefore, an absence of radiated energy for a time. Data accumulated by Jakobson, Fant, and Halle /1/ indicate that for normal continuous speech, the silence caused by a closure lasts from 30 to 100 ms. This silence is followed by a noise-like burst of sound, generally lasting

less than 50 ms for stops. Therefore, the continuant/interrupted decision is based upon the measurement of the duration of the absence of energy which is the fundamental clue in the determination of the interrupted phoneme. The normal procedure is to declare interrupted at the onset of silence, and then to declare continuant 50 ms after the beginning of noise or after 10 ms of silence. A simple sum of the energy in the frequency bands may be made and compared to the present voice level to determine if there is silence or signal.

The preceding six features are, as stated earlier, fairly well defined experimentally. However, four more distinctive features, whose acoustical correlates are not readily determined, are required to uniquely specify the 35 English phonemes. This second class of distinctive features is described in the next section, together with the knowledge concerning each.

10. Unspecified Features

10.1 Flat/Plain

Flat phonemes are characterized by a downward shift of their formants due to a reduction of the lip orifice and an increase in the length of the lip constriction. This feature is applicable to two vowels and four semi-vowels in English. The vowel /ɔ/ is flat, while the vowel /a/ is plain; the semi-vowels[1] /r/ and /w/ are flat, and the semi-vowels /l/ and /j/ are plain. As a guide, the values of the first two formants should be summed, and then this sum compared to 2000 cps. If the sum is less than 2000 cps, the phoneme is identified as flat.

10.2 Nasal/Oral

During the production of nasal sounds, acoustic energy is radiated through the nasal passages only. Consequently, the intensity of these sounds is lower than that of sounds radiated exclusively from the mouth. In English, /m/, /n/, and /ŋ/ are nasal, while /l/, /r/, /w/, and /j/ are oral. Note that there are other oral phonemes, but the earlier consonantal/nonconsonantal decision isolates the vowels from the liquids and nasals; therefore, the nasal/oral category has been reduced to those phonemes mentioned above.

There are two clues to the nasal/oral distinction which require experimental verification. It has been pointed out by Halle /27/ that the first two formants of nasals are

---

[1] See Appendix A for definition.

weaker than those of orals and that there is a formant-like (weak) energy between the first two formants. In addition, there seems to be an abrupt transition in the formants from nasals to following vowels. Finally, a zero is known to exist in the spectrum of a nasal, which might be employed, but it may be very difficult to isolate its effect. At the present, Hughes Communications Division proposes to use the output of a low-pass filter, set at 400 cps, with an experimentally determined threshold, to make this decision. That is, if the energy level out of this low-pass filter is below the experimentally determined threshold, the phoneme in question is nasal /28/.

10.3 Tense/Lax

The feature tense/lax is related to the degree of deformation of the vocal tract from its rest position. The greater (versus lesser) degree of deformation of the vocal tract from its rest position apparently leads to longer duration and greater energy of tense sounds as well as increased clarity of their formants. Fant considered the measurement of duration to be of importance for this feature /26/. His measurements show that the tense stops /p/, /t/, and /k/ have consistently greater duration than their lax counterparts /b/, /d/, and /g/. However, the feature tense/lax is not of major importance in distinguishing General American English[1] consonants since in American English tense

---

[1] See Appendix A for definition.

consonants are all voiced and lax consonants are all unvoiced. Therefore, a low-pass filter, set at 300 cps, in conjunction with a threshold detector, will suffice for this distinction for consonants.

For sonorant sounds, the genetic definition stated above may be interpreted acoustically as follows: The feature tense/lax is related to the degree to which the formants shift from their neutral position /1/. The neutral positions of the formants correspond to the resonances of the vocal tract at its rest position. Although neutral formant positions vary slightly from person to person, they are usually located at F1=500 cps, F2=1500 cps, and F3=2500 cps /26/. It is likely that these values might have to be adjusted for people who have a fundamental pitch frequency which is unusually high or low as will be pointed out later. The data of Peterson and Barney indicate that the decision tense/lax for sonorant sounds may be made in the following way: Obtain the sum of the absolute values of the deviations of the first two formants from their neutral positions and compare this to 650. If the sum exceeds 650, then the phoneme is tense; otherwise, it is lax /32/.

10.4 Strident/Mellow

The phenomenon called stridency occurs when we constrict the oral cavity, causing a rough-edged obstruction to the flow of air. When this happens, there tends to be an increase in the intensity of noise. In English, the affricates

/tʃ/ and /dʒ/ and the two labiodental fricatives /f/ and /v/ are strident, while the fricatives /θ/and /ð/ and the stops /k/ and /g/ are mellow. This feature is not required for any other English phonemes. Presently, the only clue for making this decision is to be found in the intensity of the noise produced by these sounds. Therefore, it must be determined whether strident phonemes do, in fact, exhibit greater acoustic intensity than do mellow phonemes. If this is true, then this decision will simply be implemented by a threshold detector. However, much more experimentation is necessary to determine a satisfactory measurement technique for this feature /28/.

Each feature has been described in considerable detail, and the measurement techniques presently envisioned have been delineated. To illustrate how the distinctive features may be employed to detect a phoneme, various sonagrams will be discussed.

The Sona-Graph[1] (sound spectragraph) performs a transformation on an audio waveform which puts in evidence many of the important acoustic features of speech. The output of this device is a sonagram, which is commonly a three-dimensional display of energy (darkness of line), frequency (ordinate), and time (abscissa). With certain auxiliary equipments, it is possible to get a "section" sonagram, which is a plot of energy amplitude (abscissa) versus frequency

---

[1] Manufactured by the Kay Electric Company; Pine Brook, N. J.

(ordinate) at a particular instant of time. Therefore, any organization of a signal in the time or frequency domains is visually apparent. In particular, resonances, type of vocal tract excitation, and abrupt changes in level or spectrum are readily discernible. Figure 10-1 shows ~~a~~ sonagrams of the word "faced" which includes many of the acoustic parameters listed above. Note that general spectral characteristics such as vowel resonances and the predominance of high-frequency energy in the fricative are evident. Also, it is quite easy to compute the fundamental pitch frequency by counting the number of glottal pulses (vertical striations) per unit time. The duration of the various segments and the abruptness of onset for the stop burst are also obvious. Since the dynamic range (black to white) is small (only about 20 db), the envelope shape, frequency-band energy-level ratios, and general over-all level characteristics can only be crudely estimated on the three-dimensional plot, therefore, it is helpful to study a section sonagram made for the particular instant of time of interest. The principal values of sonagraphic speech studies are to provide a qualitative indication of what kind of measurements might prove fruitful and to provide gross quantitative data on resonant frequencies, time duration, and so forth. From no other single transformation can so many important acoustic parameters of speech be viewed simultaneously.

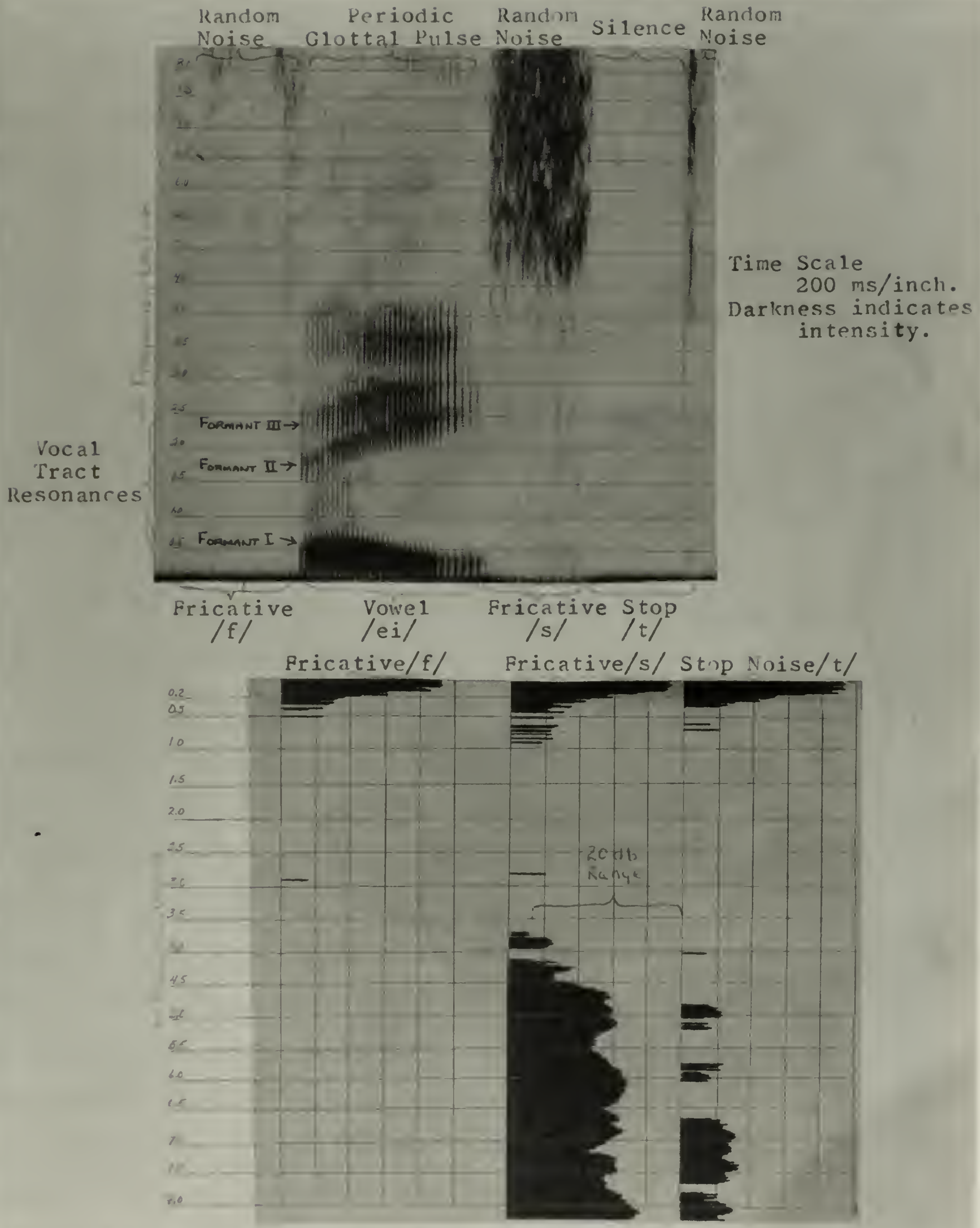


Figure 10-1. Sonagrams of the word "faced" showing various acoustic features.

11. Sonagraphic Demonstration of the Theory

Sonagrams are very useful in describing the distinctive feature analysis of phonemes. To accustom the reader to the decision procedure, the following four sets[1] of sonagrams will be analyzed:

(1) A standard CVC (consonant-vowel-consonant) utterance spoken by male speaker Number One and female speaker Number Two.

(2) The nasal /m/, as in the word "mama", spoken by male speaker Number One.

(3) The fricative /s/, as in the word "see", spoken by male speaker Number One.

(4) The stop /p/, as in the word "poppy", spoken by male speaker Number One.

The first set of sonagrams to be analyzed by the method of distinctive features is of the word "habup", pronounced /habʌp/ (See Figure 11-1,2). The CVC word /bʌp/ is preceded by the phoneme pair /ha/ so that the CVC word will receive the proper articulation. Note that there is some loss in intensity discrimination in the three-dimensional sonagram due to the multilith processing; however, it was felt that the section sonagram accompanying each three-dimensional sonagram was required anyway, and so it was decided to present the sonagrams in this way rather than in some more expensive, intensity retaining way.

---

[1] A "set" consists of a three-dimensional sonagram and a section.

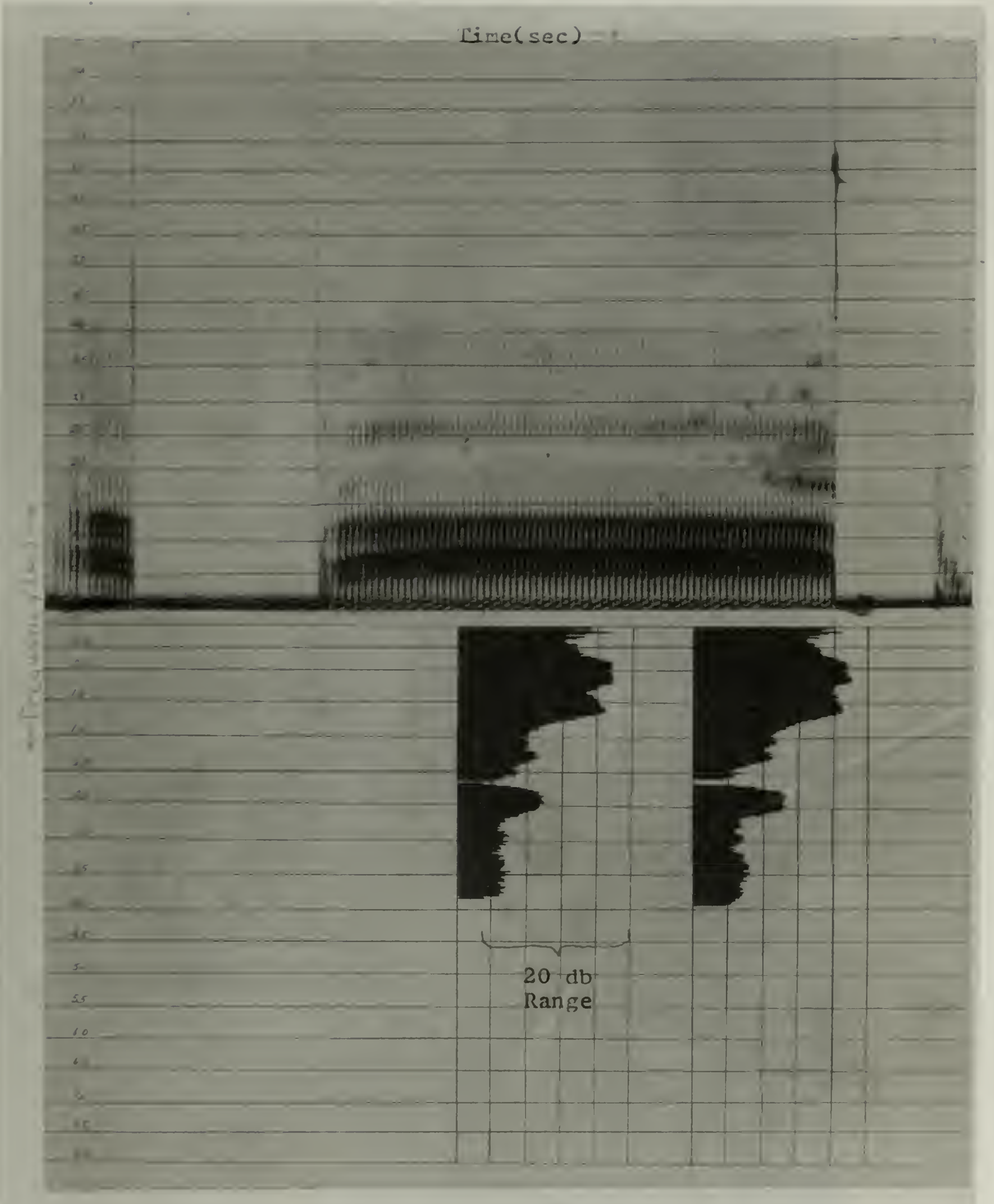


Figure 11-1. Sonagrams of the word /habup/ spoken by male speaker Number One.

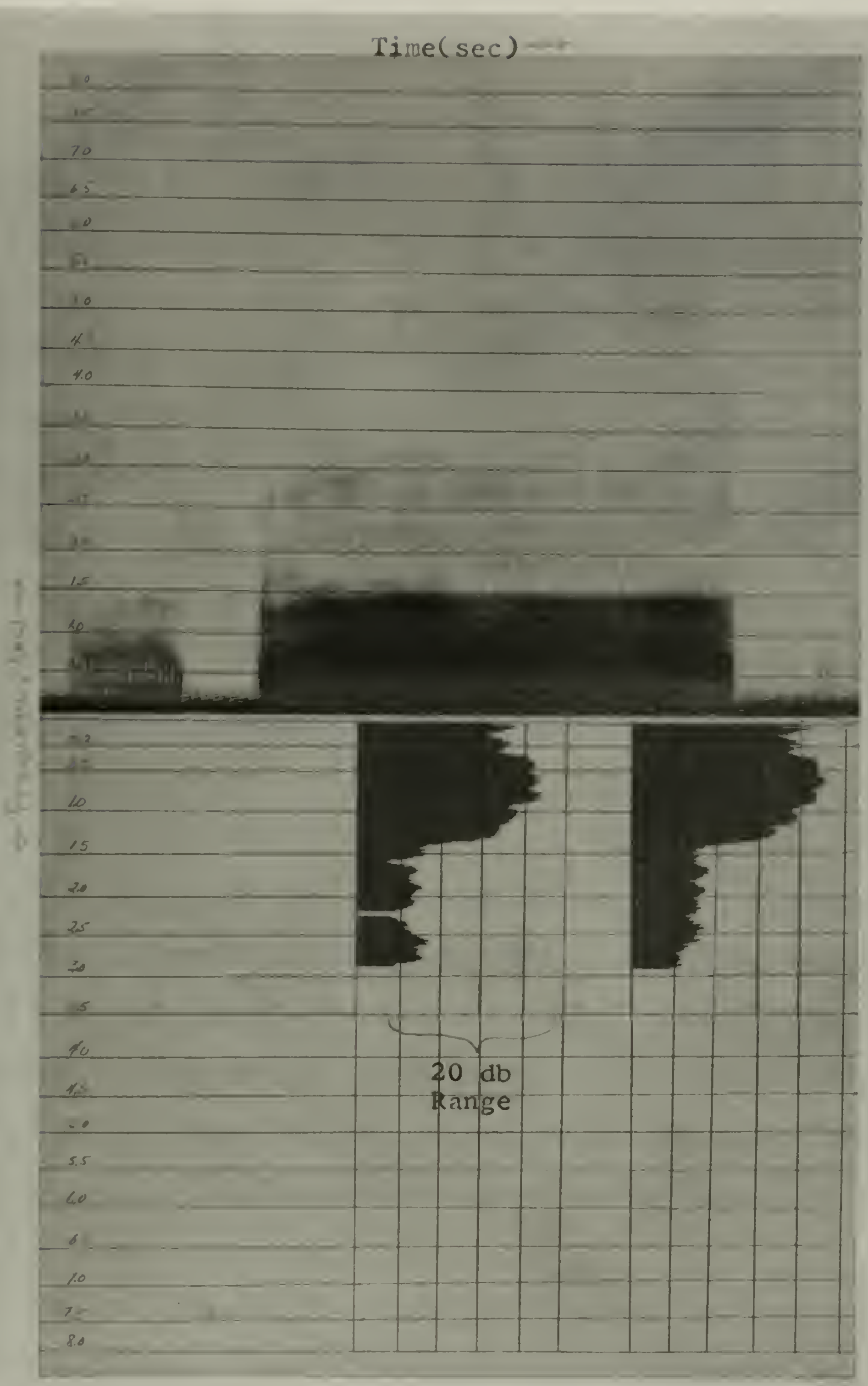


Figure 11-2. Sonagrams of the word /habuŋ/ spoken by female speaker Number Two.

Let us concentrate on the Sona-Graph representation of the phoneme /ʌ/. Referring to the tree diagram, Figure 8-2b, note that the first decision to be made is whether the sound is sonorant or nonsonorant. The acoustical correlate of this feature, as related by Hughes, is determined by comparing the energy of the sound in the band 300 cps to 1 kc with the energy in the band above 2.5 kc /25/. If the energy in the lower band is greater, then the sound is sonorant. From Figures 11-1 and 11-2, the sections of the male and female utterances, it may be determined that the energy in the lower band is indeed greater. The area ratio is about two to one for the male speaker and slightly more for the female. Therefore, the sound is sonorant.

The next decision indicated is whether the sound is consonantal or nonconsonantal. The transitions between the sounds are of importance in this decision and it is obvious from the sonagrams that there is a transition to greater energy at the onset of the utterance of /ʌ/ and a transition to less energy at its termination. Accordingly then, the sound is nonconsonantal.

The grave/acute decision is based on the amount of separation between the first and second formants. Upon examination of both the sonagrams for the male and female speakers, the following data may be extracted:

	Male	Female
First formant (F1)	550 cps	600 cps
Second formant (F2)	1100 cps	1000 cps
F2 - F1	550 cps	400 cps

Since the differences are less than 850 cps, the sound is grave. Having determined that the first formants are both greater than 500 cps, we can also state that the sound is non-diffuse.

Finally, the tense/lax decision is determined by adding the absolute differences between the first formant and 500 cps and the second formant and 1500 cps. In this case, the sum is 450 cps for the male speaker and 600 cps for the female speaker. Therefore, the sound is lax and is identified as the phoneme /ʌ/.

The next set of sonagrams to be analyzed is shown in Figure 11-3. Male speaker Number One has recorded the word "mama" and it will now be shown how to extract the nasal phoneme /m/ by the "method of distinctive features". Note that the first and third sections are of the phoneme in question.

As before, the sonorant/nonsonorant decision must be made first. The third section indicates that the frequency band from 300 cps to 1 kc contains more energy than the band above 2.5 kc by a ratio of about 4.5 to 2.5. The three-dimensional sonagram also indicates clearly that the sound in question is sonorant.

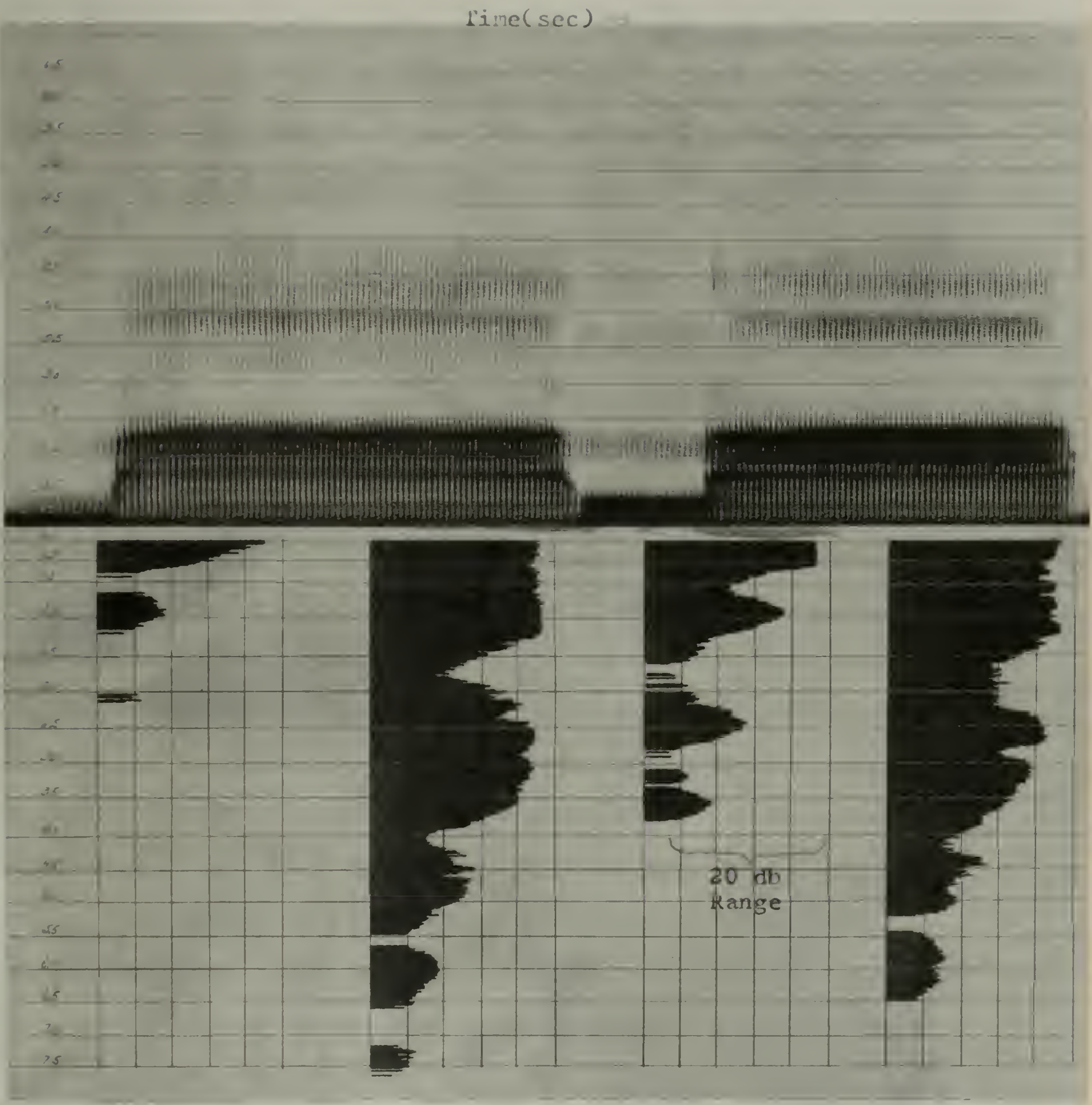


Figure 11-3. Sonagrams of the word /mama/ spoken by male speaker Number One.

The tree diagram indicates that we must next decide whether the sound is consonantal or not. Again the energy transitions are important and it is clear from the sonagrams that there is a drop in radiated energy during the utterance of /m/ and therefore, it is consonantal.

We must next make the nasal/oral distinction and the characteristic phenomena evidenced by the sonagrams are very interesting. Primarily, note the abrupt transitions of the formants between the phoneme /a/ and the nasal /m/. Also, while the formants of the /m/ are very distinct, they are considerably lower in energy than those of the phoneme /a/. Finally, weak formant-like energy between the formants is clearly indicated. While it is true that these features are visually evident on the sonagram, they are fairly difficult to extract in a computer; therefore, Hughes Communications Division is planning to use a low-pass filter whose cut-off frequency is set at 400 cps to determine this feature. That is, if the output of this filter is below an experimental threshold, then the utterance is nasal. The feasibility of this method is apparent when you compare the 0-400 cps band on the section sonagrams of the nasal /m/ and the oral /a/.

The acoustical correlate of the compact/noncompact decision is based on the location of the first formant. Upon examination of the sonagram set, it appears as if the first formant is located just under 1 kc. However, a closer inspection will indicate that voicing and the first formant

have combined to "peg" the Sona-Graph in the region from 100 cps to 300 cps on the third section. This could have been more clearly indicated by dropping the mark level on the sonagram; however, the upper formants would have been lost and it was felt that it was more important to show these as their composition and placement are characteristic of nasals. The first formant is therefore less than 600 cps and the sound is noncompact.

The second formant is located at about 900 cps and it was previously stated that the first formant was at about 250 cps. The difference between the two is 650 cps and this is less than 850 cps, the threshold as determined by Peterson and Barney; therefore, the sound is grave and is identified as the nasal phoneme /m/ /32/.

Hughes' work /25/ is evidence of the feasibility of the process just outlined for vowels and nasals. However, fricatives and stops are slightly more difficult to discern due to their shorter time duration and lower overall energy. The method is applicable at any rate, and does show considerable promise as evidenced by the following discussion.

Figure 11-4 contains the set of sonagrams for the phoneme pair /si/, pronounced like the word "see". Let us apply the "method of distinctive features" to the phoneme /s/, one of the more common fricatives.

The sonorant/nonsonorant decision is trivial in this case as there is no apparent energy in the lower band, 300 cps to

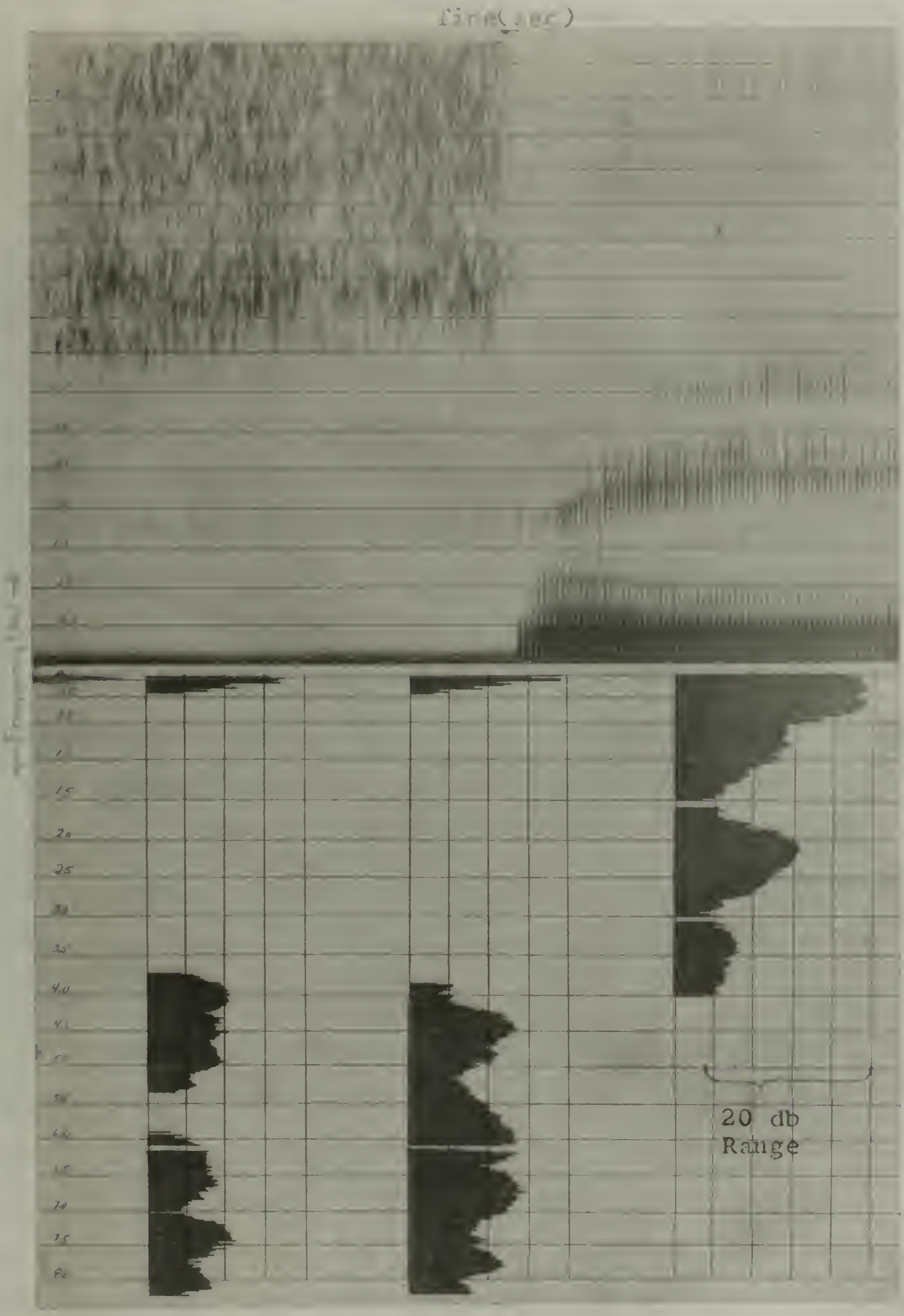


Figure 11-4. Sonagrams of the word /see/ spoken by male speaker Number One.

1 kc, and a considerable amount in the upper band above 2.5 kc. Therefore, the sound is nonsonorant.

The consonantal/nonconsonantal decision is somewhat harder to make based upon the transitions evident on the sonagrams. However, close calculation will indicate that there is indeed a transition to a greater amount of radiated energy as the phoneme /i/ was uttered, and hence, the sound is consonantal.

The tree diagram shows that the next decision to be made is whether the sound in question is continuant or interrupted. In the production of the word "see", the vocal tract is obviously not completely closed at any time as evidenced by the three-dimensional sonagram. That is, there is no onset of silence at any time, and therefore, the sound is continuant.

The feature tense/lax is easily determined in American English due to the fact that tense consonants are all voiced and lax consonants are all unvoiced. The sections clearly indicate that there is voicing present, and hence the sound is tense.

The acoustical correlate of the feature compact/non-compact is determined by comparing the energy in the band 1.7 kc to 3.4 kc with the energy in the band 700 cps to 1.4 kc. This is to show whether or not sharp resonances occur in the band 1.7 kc to 3.4 kc. In this case, the results are again quite obvious in that there is no appreciable energy in either band. The sound is consequently classified as non-compact.

The feature grave/acute, as applied to fricatives and affricates, involves a comparison of the energy in the band 4 kc to 7 kc with the energy present in the band 700 cps to 7 kc. Acuity is evidenced by a concentration of the radiated energy in the higher frequencies of the audio spectrum. The section sonagrams clearly indicate that practically all of the energy present in the sound is in the higher frequencies and hence, the sound is acute and identified.

As a final example of this procedure, let us attempt the identification of the stop /p/ as in the word "poppy" or phonemically, /papi/. Figure 11-5 is the set of sonagrams representing this word.

The sonorant/nonsonorant decision is quite easily made in this case as there is a considerable concentration of energy in the lower band from 300 cps to 1 kc. This large amount of energy is due to the fact that male speaker Number One has an unusually low fundamental pitch frequency and emphasized the phoneme /p/ so that it's characteristics would be more distinct. At any rate, the sound is sonorant.

The feature consonantal/nonconsonantal is not as easily rationalized in this case for the reason just mentioned. That is, the over-emphasis of the phoneme /p/caused an unusual rise in the energy level of the random noise at the end of the utterance. However, it must be remembered that the sound /p/ beings with a closure of the lips (indicated by the period of silence on the sonagram) following the phoneme pair /pa/

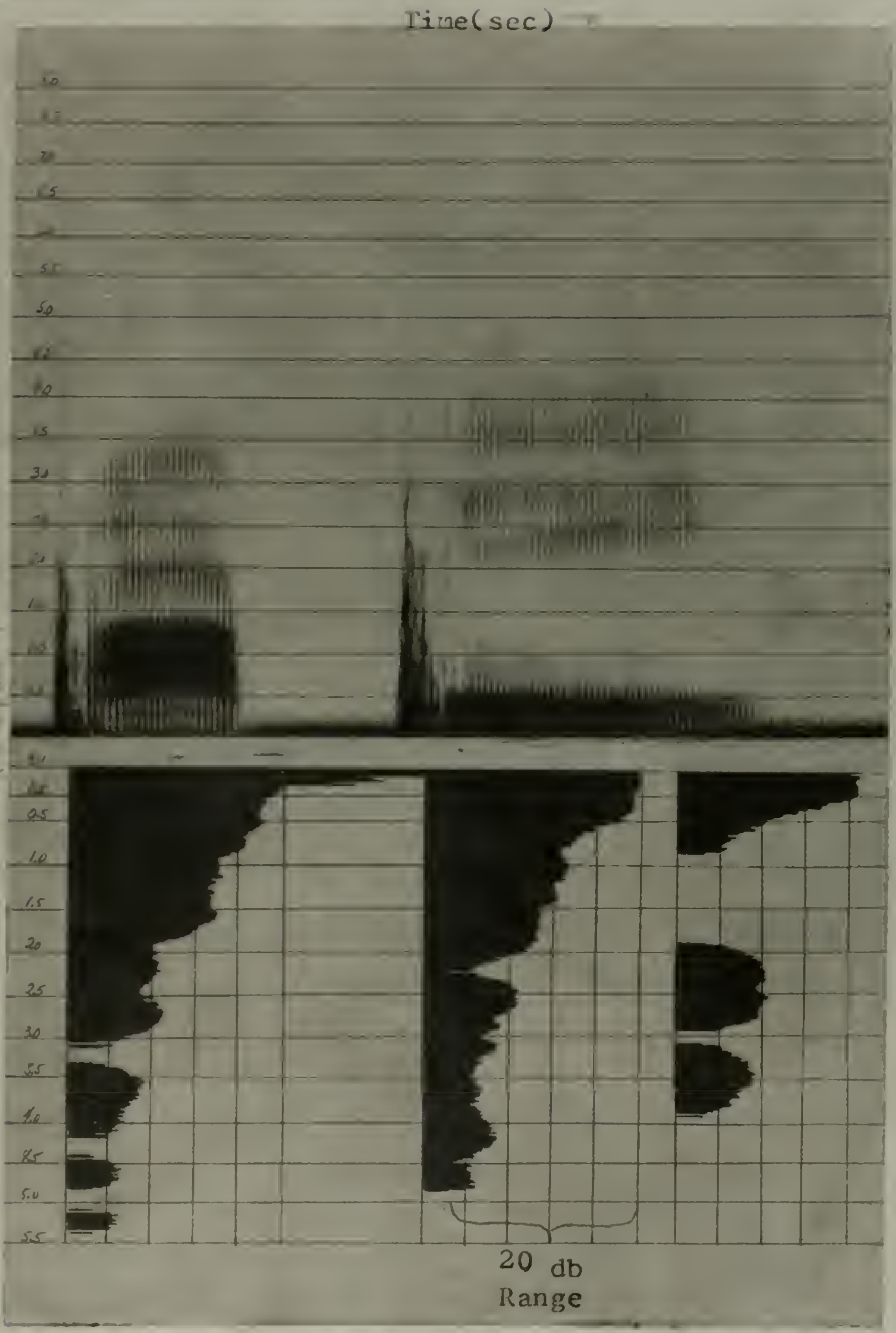


Figure 11-5. Sonagrams of the word /poppy/ spoken by male speaker Number One.

and hence, there truly is a drop in overall energy, and the phoneme is consonantal.

As stated above, the phoneme consists of a period of comparative silence followed by a burst of random noise. Therefore, the phoneme is classified as interrupted.

In order to determine whether or not this stop is tense or lax, note that even during the silent part of the utterance, some voicing is evident. During the burst part, there is considerable voicing apparent, and hence the sound is tense.

The determination of compactness, in this case, involves a relative comparison of the energy ratio of two bands of the /p/ spectrum with the same band ratio of the back variants of the /k/ or /g/ spectrum. The data of Halle, Hughes, and Radley /31/ indicate that the back variant of /g/ has about 40 per cent of its energy in the 1 kc to 2 kc band. The section of the sound in question, contains something less than 30 per cent in this same range. Therefore, the utterance is noncompact.

The grave/acute decision is elementary in this case as there is obviously a heavy concentration of energy in the lower frequencies. Referring to the section, it is quickly determined that the energy in the band above 2.5 kc is about one third the energy in the band above 700 cps. Hence, we have identified the phoneme as the stop /p/.

The "method of distinctive features" has been defined and typical examples of its application described. To develop

more rigid measurement procedures, additional research is needed which will undoubtedly involve the processing of a voluminous amount of data. The modern high-speed digital computer is especially adapted to the reduction of large amounts of data, and consequently is one of the major tools in speech research today. The United States Naval Postgraduate School is in a unique position in this respect due to its fine computer facility. However, a speech input device will be necessary to reduce the acoustic waveform into a quantized time-amplitude-frequency spectrum for computer manipulation. Such a device is described in detail in the next section.

12. Design of the Acoustical Input Device

The basic design of this acoustical input device is modeled after the standard fixed-channel vocoder. A two-channel tape loop feeds a set of contiguous bandpass filters, each with its own amplifier, rectifier, and smoother. All channels are then multiplexed and fed to an analog to digital converter and into the computers.

Specifically, the input speech will be recorded on one channel of a two-channel tape loop and the timing for the SDS[1] 11-bit analog to digital converter recorded on the other channel so that samples of the speech will be taken at precisely the same instant. This is necessary because the multiplexer now on order for the school is an SDS 16-channel unit and therefore, the speech sample will have to be applied twice as the number of proposed channels is 32. Considering the sampling rate, estimated program length, and the size of the CDC[2] 1604 memory, it is felt that at first, the speech sample should be limited to one second. This is not as restrictive as it sounds, as one sonagram represents 2.4 seconds of speech and quite a bit of speech may be analyzed in one sonagram. A block diagram of the complete input device is shown in Figure 12-1.

It was decided that 32 channels would be adequate for this installation. This decision is based upon the results

---

[1] Scientific Data Systems; Santa Monica, California.

[2] Control Data Corporation; Minneapolis, Minnesota.

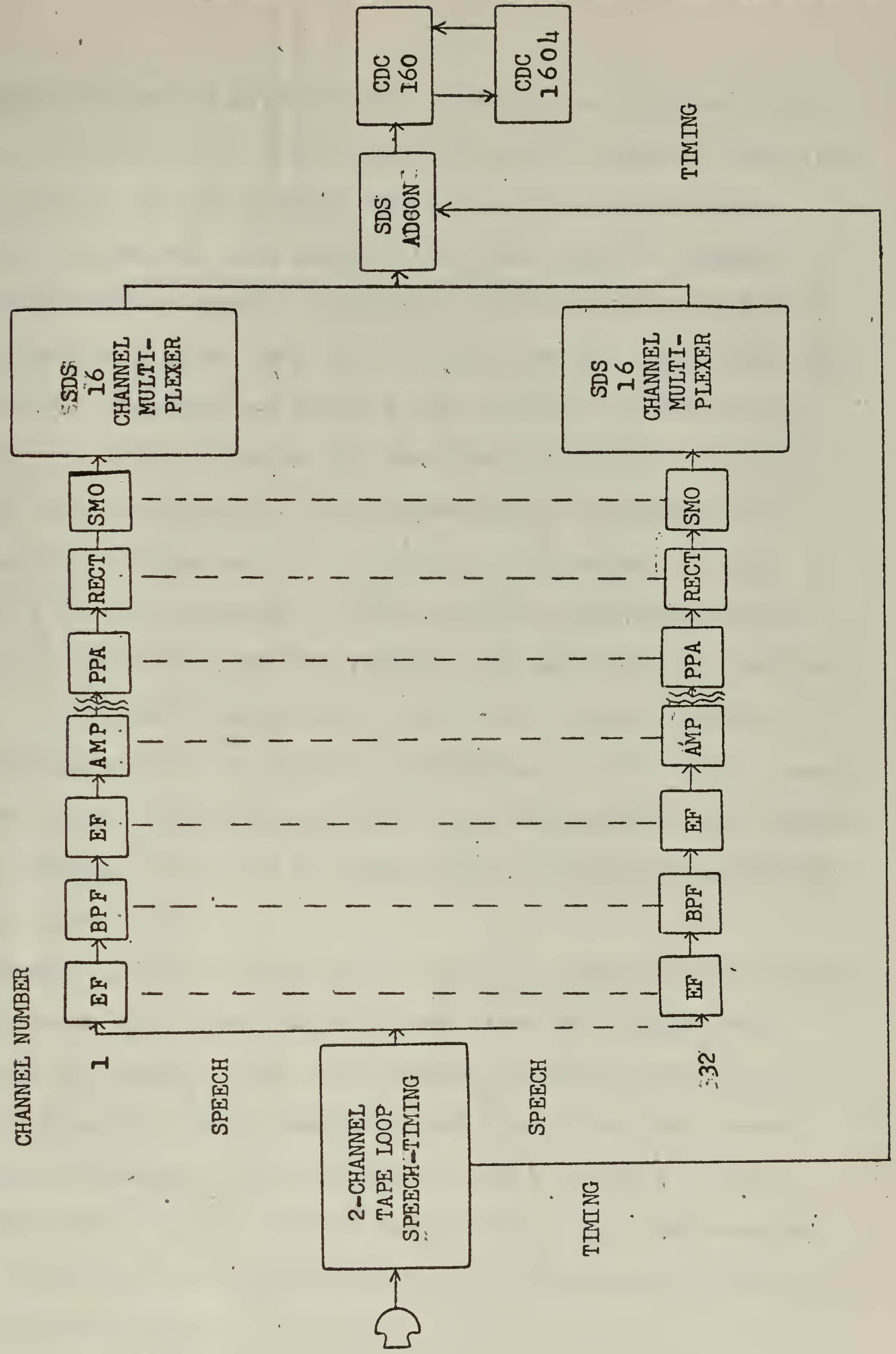


Figure 12-1. Block diagram of the proposed acoustical input device.

of similar research accomplished by Hughes at Purdue /25/. Hughes obtained quite good results with 35 channels and after consultation, it was decided that with the postgraduate school's equipment, the benefits derived from 32 channels far exceeded the losses. Channels 15 and below should be de-emphasized at the rate of 3 db per octave. This will aid the formant tracking portion of the program to distinguish between the first formant and the first or second harmonic of the voicing frequency. The lower order components of voicing for females are particularly troublesome. Also, channels in the vicinity of 1800 cps are pre-emphasized to lower the confusion between formant two and formants one and three. The filter bandwidths, center and end frequencies, and weightings are indicated in Appendix C. The filters must be very sharp, with slopes of at least 50 decibels per octave on the skirts from the 3 db down point, with no "bounceback" greater than 50 db.

Figure 12-2 is the complete circuit design for one channel. All circuits have been tested except for the transformer coupling indicated. The transformers selected must give the proper input and output impedance and also have the correct frequency response. This design is quite similar to the equipment built by the author earlier this year and many pertinent problems of construction are discussed in the report of that project /33/.

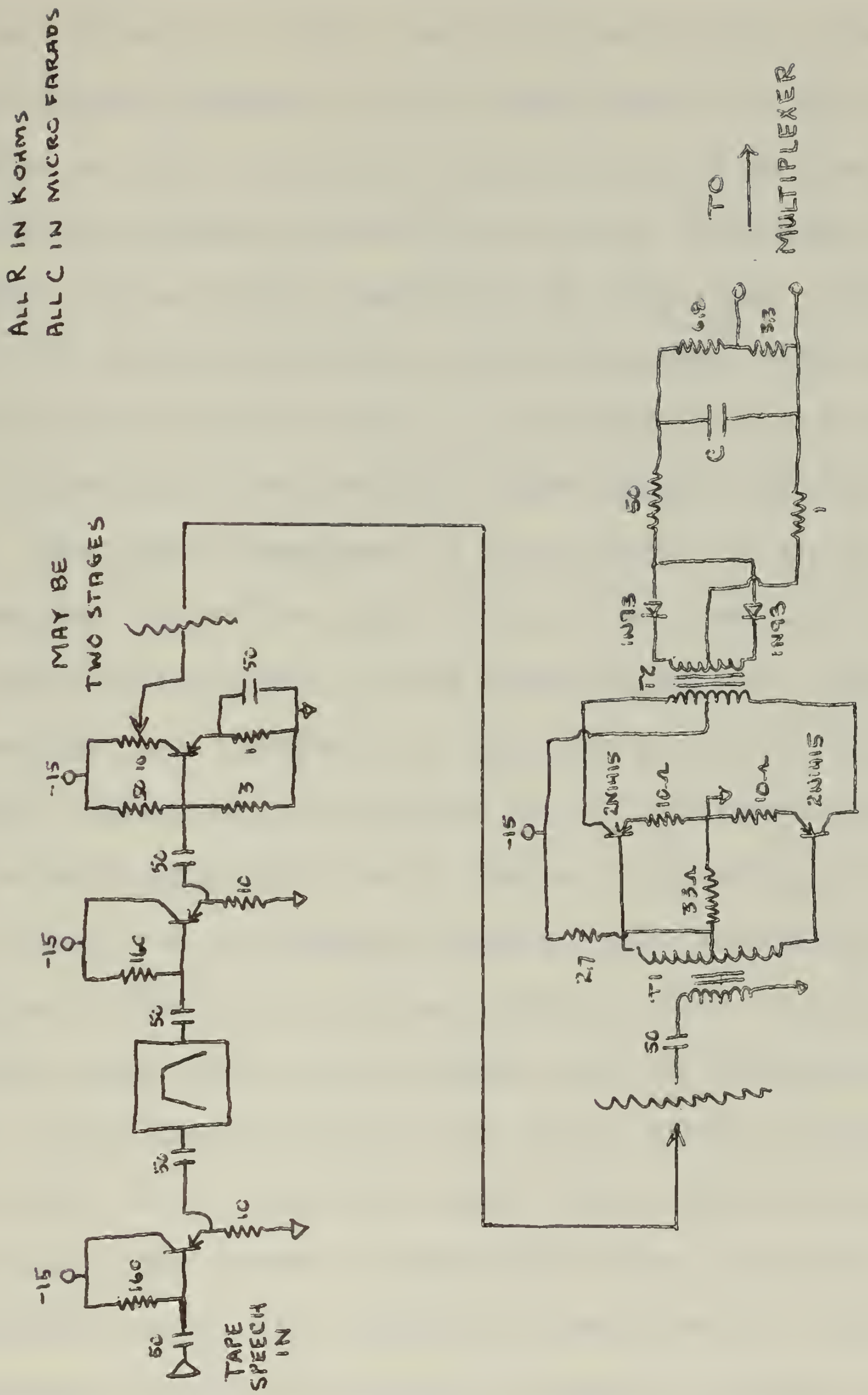


Figure 12-2. Circuit diagram for one channel of the proposed acoustical input device for the Digital Control Laboratory, U. S. Naval Postgraduate School.

The first stage in each channel is of course an emitter-follower circuit so that the input device will have the proper input impedance for the two-channel tape recorder, and also so that the filters will be isolated from each other. The emitter-follower shown has an input impedance of 110 kilohms, and an output impedance of 2000 ohms. Since there are 32 of these in parallel, the equivalent input impedance of the device will be about 3.5 kilohms which is high enough not to load the tape recorder whose output impedance is 600 ohms. The input impedance of the filters is at least 10 kilohms and therefore, they do not load down the followers.

Emitter-followers of the same design also follow the filters so that they will be isolated from the succeeding stages. Following this second set of emitter-followers are simple audio amplifiers each having a maximum gain of about ten. More than one stage of audio amplification may be required to obtain the proper voltage level for the multiplexer, but this will have to be decided when the complete circuit is built. Note that up until this point 2n526 transistors have been used. It is important that these transistors or equivalent substitutes be used since they are good, low noise, transistors especially adapted to audio work. The input impedance to these amplifiers is about 57 kilohms which does not alter the 2000 ohm output impedance of the emitter-followers. The output impedance of the amplifiers is about 12 kilohms and a suitable audio (20-10000 cps) transformer must

be selected to match this impedance.

The next stage is a Class B, push-pull amplifier. This stage is needed so that a good linear waveform will be presented to the following full-wave rectifying circuit. The voltage divider circuit composed of the 33 ohm resistor and the 2700 ohm resistor gives a slight forward bias on the transistors to prevent cross-over distortion. The 10 ohm emitter resistors are for stabilization. Note that it is obviously important to maintain a common ground throughout each channel, eliminating the possibility of a full-wave bridge rectifier directly following the audio amplifier stage.

The final stage is a full-wave rectifier stage followed by a smoother whose time constant may be varied by adjusting the value of the capacitance. To retain all of the information conceivably available in each channel, it is important that accurate envelope detection be made. It is believed by Denes /34/ and others that the envelope may have relatively high frequency[1] variations during transitions and that it might be important to be able to ascertain when these variations occur. Having established the desirability of retaining all the envelope information available, it is interesting to discuss the various factors which affect the amount of information retained. These factors are:

---

[1] The envelope of the fully rectified wave may contain frequency components up to 400 cps.

(1) Bandwidth of the filters;

(2) Smoothing time constant;

(3) Smoothing ripple that will be tolerated;

(4) Sampling rate of the analog to digital converter.

As a crude example of the influence of these factors, let us consider a waveform composed of a carrier tone at 5130 cps, modulated by a signal which has numerous frequency components. This carrier happens to be the center frequency of filter number 28 whose bandwidth is 700 cps. We know that due to the bandwidth of the filter, no modulation frequency higher than 700 cps will be passed. In addition, the smoothing time constant of the circuit limits the highest modulation frequency to be passed. For example, if the time constant is too long, the envelope will not be accurately detected; the detected waveform that results will only contain frequency components to the frequency at which the smoothed waveform fails to follow the peaks of the rectified wave. Therefore, to obtain all the information in the envelope that gets through the filters, it is necessary to adjust the time constant of the smoother so that the smoothed waveform follows the peaks of the fully rectified wave. Obviously we cannot do this exactly as we have only decay between peaks; however, we can adjust the time constant to limit the decay to some percentage of the peak value obtained on the previous pulse. Hughes' work indicates that a negligible amount of information is lost if you allow ten per cent decay per half

cycle [25]. This arrangment will generate a certain amount of ripple, but this is also negligible if the decay is limited to ten per cent. Finally, if the sampling rate is not the Nyquist rate for the highest frequency passed, then there will be an additional loss of information. Now, the speech waveform is much more complicated than a single modulated carrier; however, the same principles apply. Therefore, it was decided to adjust each smoother for its particular center frequency and to allow ten per cent ripple. No consideration will be given to the effect of the filter bandwidths on the sampling rate, as it is convenient to sample all channels at the same rate to maintain the form of the data. The time constant of each smoother is then given by the following equation; $t_c = C \times 10^4 = 1/.2f$. From this equation, the value of the capacitance for each smoother can be calculated, given a particular center frequency. The sampling rate should be such that at least 400 cps information is maintained.

There are 16 channels to be sampled on each pass, and if we sample each at the Nyquist rate of 800 samples per second, we would require analog to digital conversion at the rate of 12,800 times a second. The particular converter on order for the school operates at up to 20,000 times a second and therefore is quite adequate for this installation.

As the data is introduced into the computer, it will have to be manipulated into the proper form and then processed appropriately to complete the identification. The next section describes the form of the data and indicates in flow-graph form, the required programming to process the data.

13. Form of the Data and Required Programs

With the speech sample recorded on one channel of the two-channel tape loop, and all gain levels adjusted so that the signal range throughout the ten kilocycle spectrum is within the range of the multiplex unit, the data may be read into the computer. The equipment is such that four samples may be packed into one cell of the CDC 1604. The memory space required will be 3200 cells per 16 channels for each one-second speech sample. Therefore, a total of 6400 cells will be required for all the data. A simple index-mask-shift routine will provide easy access to the data.

It is interesting to reflect briefly on just exactly what the extracted data represent. A careful study of the circuitry will indicate that a sample will simply be the amplitude of the modulating envelope for a particular instant of time. This amplitude picking, in conjunction with the smoothing indicated, gives an output which is in one-to-one correspondence with the power in the band during the sample time. That is to say, each sample is in one-to-one correspondence with the energy in the band at the instant of time in question. Mathematically it turns out that the difference is due to a constant, but nevertheless, as the energy varies, so do the amplitude data and this is the desired result. Therefore, it is sufficient to say that the addition of the outputs of several contiguous filters will give a sum which is indicative of the "energy" in the band covered by

these filters.

In the first phases of the research, it was decided that every four samples (of 1.25 ms each) should be averaged together prior to processing. While this degrades the information content somewhat, it makes the initial program checkout much easier as there is less likely to be any abrupt changes in the data. In addition, the development of a good formant tracking program is required prior to any further work, and research indicates that changes occuring within 5 ms are not of importance in tracking the formants of vowels, nasals, and liquids. It is further recommended that each averaged set of data be looked at independently at first, and then an effort should be made to extend the influence of one set on its adjacent sets.

The flow-charts described in the following paragraphs depend to a great extent on the determination of certain thresholds which are dependent on the particular equipment used. These thresholds may be calculated by observing the data derived for a carefully selected group of phonemes. For example, if it is desired to determine the threshold required for the sonorant/nonsonorant decision, the energy distribution of the following "high" vowels;/i/, /I/, /u/, and /U/, which are sonorant, should be compared with the energy distribution of the voiced fricatives /θ/, /ð/, /v/, and /z/, which are nonsonorant. In other words, data should be taken for the "borderline" cases, and the thresholds defined on the basis

of this information.

A "building block" approach, which closely follows the tree diagram given earlier, must be used in the programming of this problem. This is due to the fact that we do not yet know how to obtain all of the features by this method of analysis. However, we do know how to obtain certain parameters from the data (e.g. formant frequencies, energy levels, etc.), and having once reduced these methods to programs, the search for the acoustical correlates of the unspecified features may begin. Therefore, the required programming logically breaks down into the following subroutines:

(1) Formant tracking;

(2) Smoother;

(3) Boundary;

(4) Classification.

The requirements for these programs will be described in detail in the following paragraphs of this section.

A good formant tracking routine is of obvious importance. Many of the decisions required depend on the accurate determination of the formant frequencies and their transitions. The formant tracking subroutine flow-graphed in Figure 13-1 depends upon a very simple and direct relationship between the formant frequencies and the short-time spectral data in the computer. That is, the filter channel whose center frequency is closest to the actual formant frequency at the time the filter output is sampled will be a maximum with

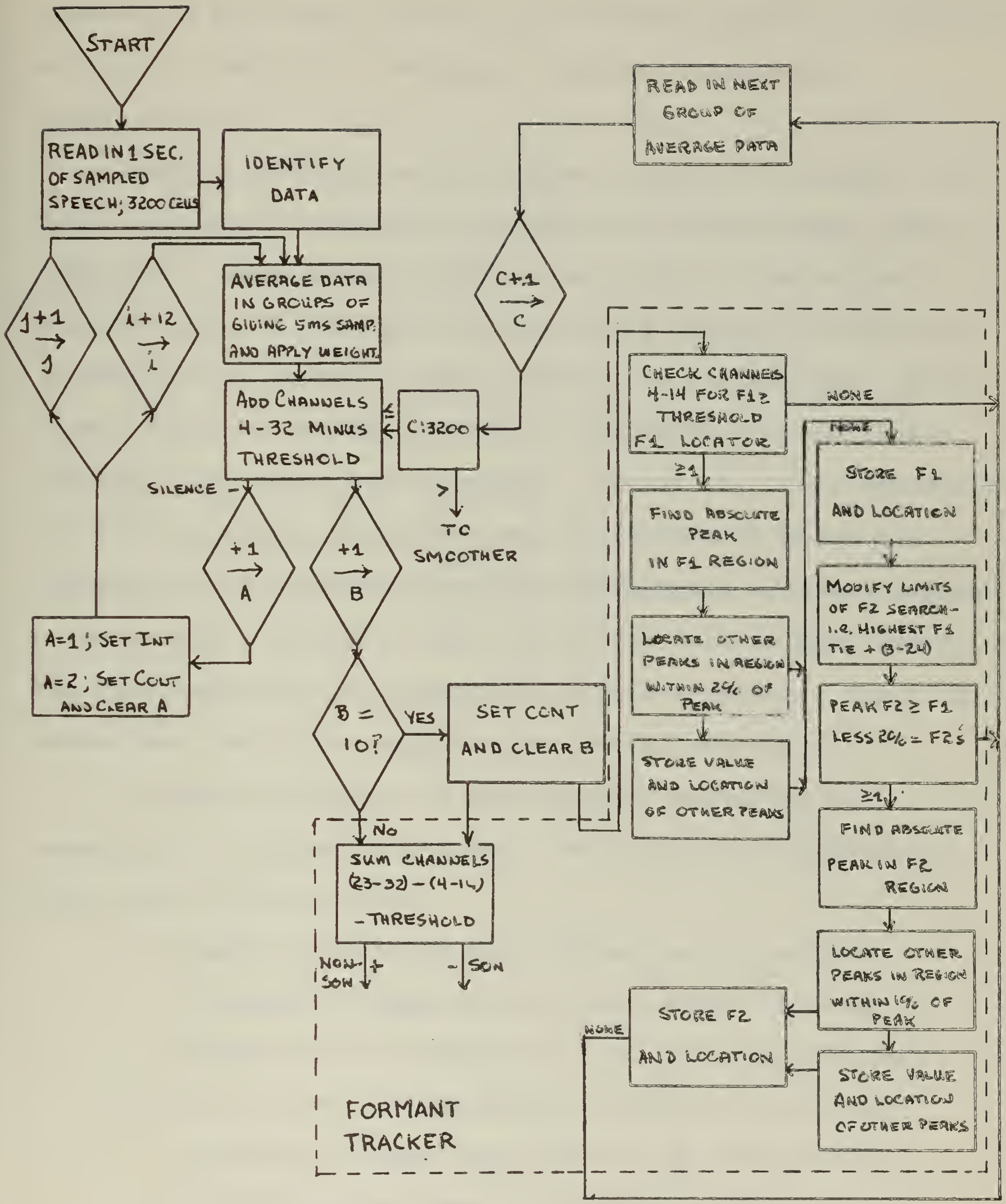


Figure 13-1. The formant tracking subroutine with continuant/interrupt lead-in.

respect to the other filters in the formant region. Therefore, we have a simple peak selecting routine with certain modifications.

These modifications are a direct result of the fact that data collected on formant frequencies of the male and female speakers of English show that F1 may occur in the region 250-1200 cps, F2 from 600-3000 cps and F3 from 1700-4000 cps. In addition to formant region overlap, it is possible that a strong first or second harmonic of a high-pitched voicing frequency may enter the F1 region. Therefore, it is impossible to define mutually exclusive sets of filters in which the maxima will be related to particular formants with any degree of certainty. Finally, while it is true that in most cases, the amplitudes of the formants drop with frequency, there are exceptions due to the vagaries of glottal spectra.

In order to retain the peak-picking technique so adaptable to a fixed filter set, the following modifications were made to this scheme:

(1) Heavy attenuation was applied to the voicing frequency region so as to reduce the relative effect of the harmonics of the voicing frequency;

(2) The entire F1 region was searched for the lowest frequency formant and then the F2 region was modified based upon the location of F1;

(3) Ties were allowed and recorded if their amplitudes were within a certain percentage of the maximum

amplitude determined for the particular formant in question.

The smoother subroutine is shown in Figure 13-2 and is necessary for phonemes exhibiting formant construction (vowels, nasals and liquids). This routine is simply for the purpose of imposing continuity of formant position with time. This was accomplished in the following way:

(1) If there are no ties, exit to boundary subroutine;

(2) If there is a tie between adjacent channels, take the average frequency as the location of the particular formant in question. If there is a tie between widely spread channels, select the one which is closest to the location determined for the same formant on the previous sample;

(3) As a final step in the smoother subroutine, excessive jumps in F1 and F2 between adjacent samples should be smoothed in accordance with Figure 13-3. This is necessary because sometimes a formant will weaken considerably in intensity for a short period (due to changes in the glottal spectrum), causing the simple peak-picking routine to find a second order maximum in the region or perhaps track an adjacent formant until the true formant returns to normal amplitude. Also, during the utterance of vowels whose formants are closely spaced, (e.g. /a/ or /i/), F1 does not always exceed F2 in

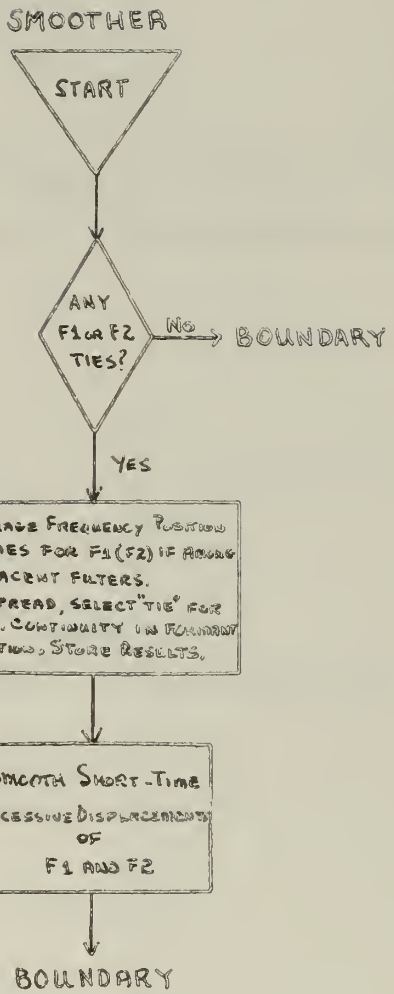
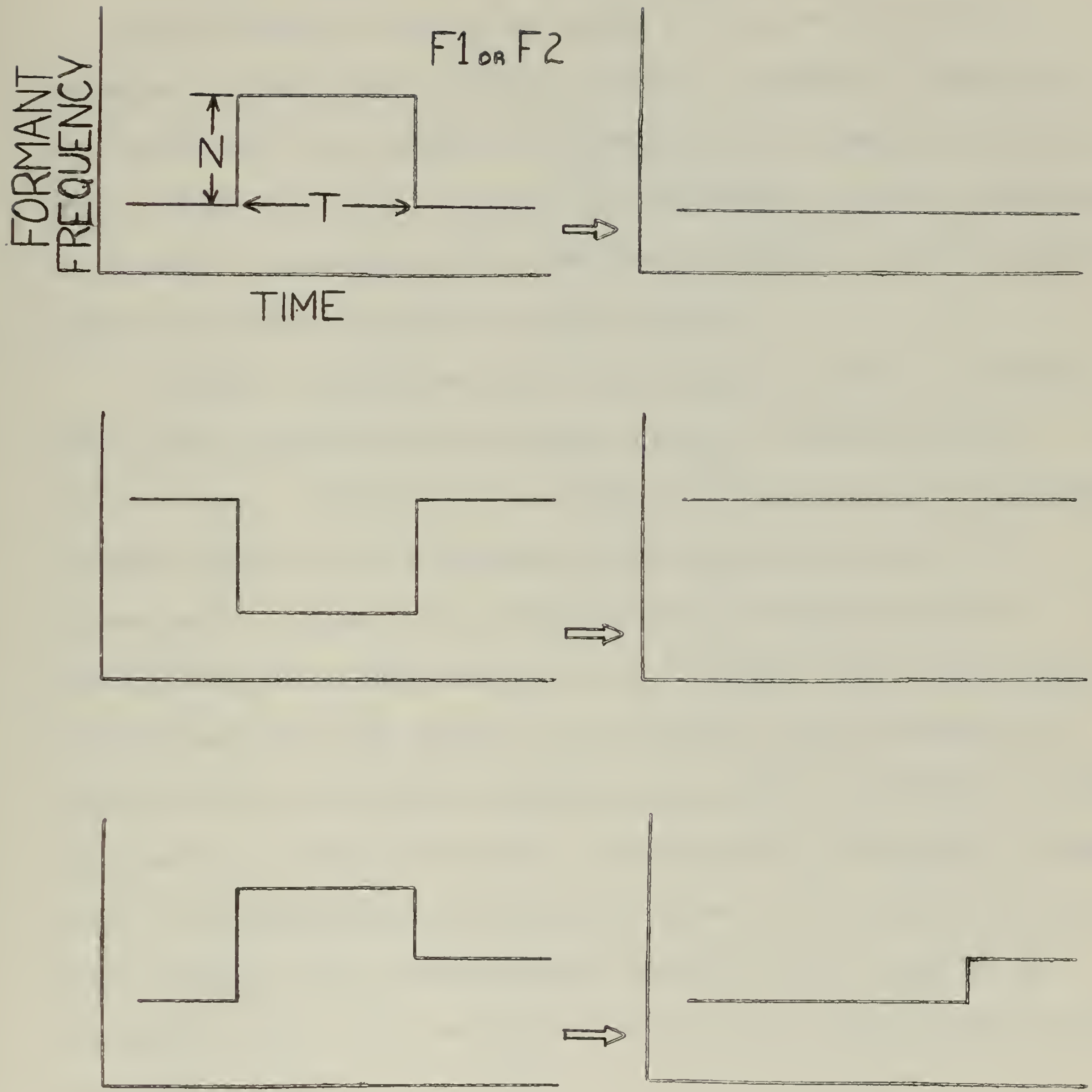


Figure 13-2. The formant smoother subroutine.

[illegible]



N = 3 filters

T = 20 sec. (4 averaged samples)

Figure 13-3. Smoothing criteria for excessive displacements.

[illegible]

amplitude, and again a simple peak picking scheme may jump randomly between adjacent formants.

The boundary routine is shown in Figure 13-4 and is designed to mark the boundary between phonemes. Boundaries are indicated by a change in formants or a change in level or both. Therefore, the routine is dependent upon the resolution of certain thresholds which may be determined in the general manner indicated earlier in this report.

Finally, all of the short routines required to extract particular features are grouped into the classification subroutines. Figure 13-1 contains the continuant/interrupted routine preceding the formant tracking routine as it is convenient to make the continuant/interrupted decision immediately upon entrance into the program. No classification subroutines will be given in this paper as the methods of obtaining the specified features are clearly outlined in the discussion of those features. In addition, all that is known about the unspecified features is given in the discussion of those features and considerable research will have to be performed before the methods of extracting these features may be delineated.

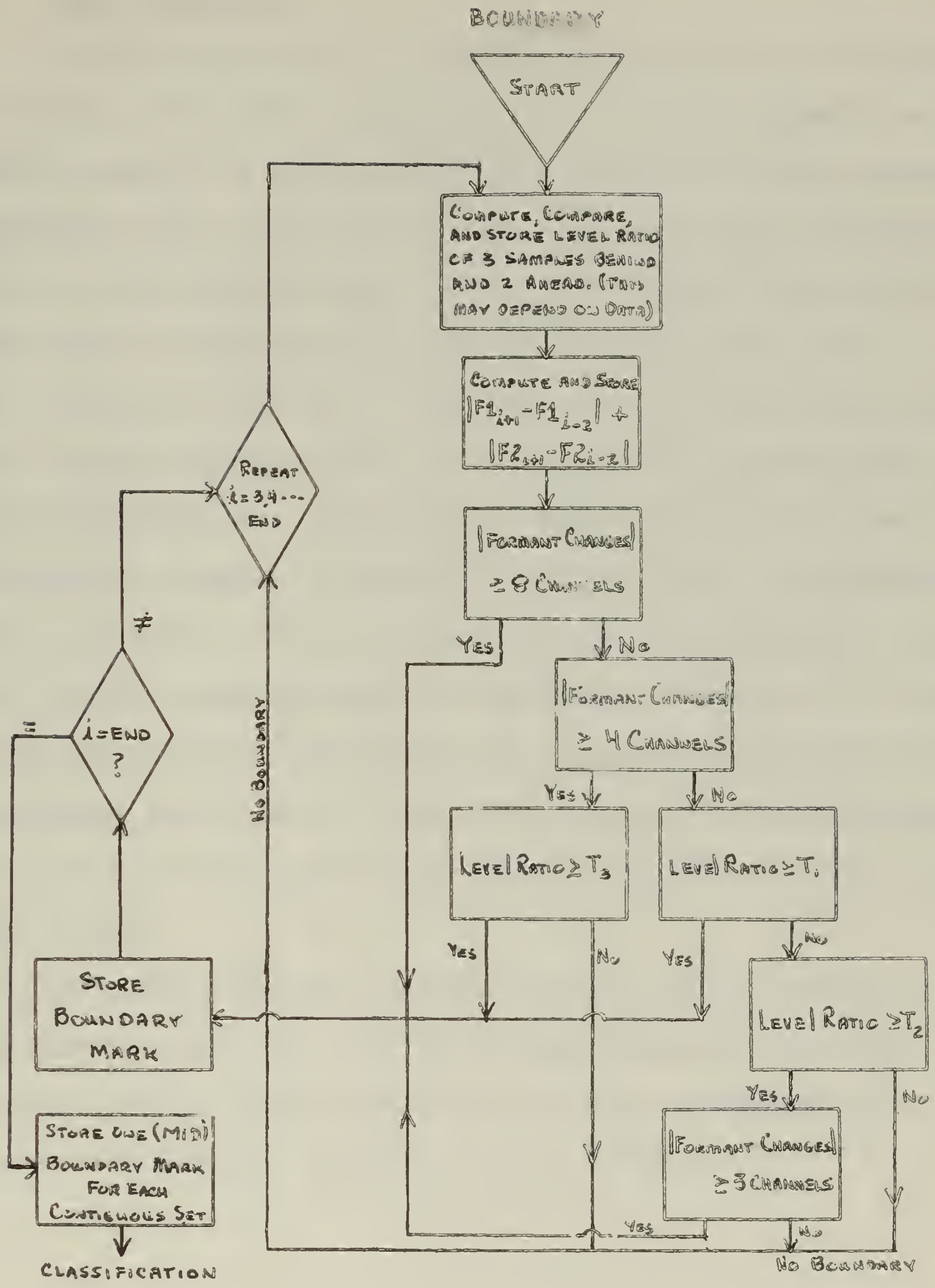


Figure 13-4. The boundary subroutine.

[illegible]

14. Test Procedure

For the purposes of threshold determination, program check-out, and system evaluation, the initial speech samples should consist of simple CVC words spoken into the recorder microphone with the mouth at a distance of about one foot from the microphone head. Later, meaningful words may be used such as those used by Hughes in his project /25/.

The microphone and two-channel tape recorder used for this research should have a flat frequency response over the range 20-10000 cps. The speech samples should be taken in an anechoic chamber so that little noise will be introduced into the system. It is important that an appropriate input level at the tape recorder be maintained (about + 1 on the volume unit meter) throughout the research program as the thresholds incorporated into the program will be determined for speech at that level and they will not be correct for any other level.

The timing should be placed on the second channel of the two-channel tape loop after the speech samples have been taken, as there is less likelihood of confusing the speakers if they are not required to speak at a certain time.

15. Conclusions

This paper provides a summary of the state-of-the-art of important speech compression devices. As pointed out, their basic faults are due to the fact that their underlying theories do not consider the speech signal to be a particular signal with certain unique properties. This fact prohibits these devices from realizing the theoretical limit of compression.

For broad tactical military use, bandwidth compression, intelligibility and quality are of primary interest. The "method of distinctive features"offers a fresh, linguistically sound approach to the problem of speech recognition, the primary process in any speech compression device. The many advantages of this approach, such as simplicity, guaranteed completeness of final solution, and consistency with observed human behavior in perceiving speech, furnish strong enough evidence for basing analysis schemes on this theory. The measurements described for extracting the specified features are not complicated and conceivably could be done with equipment of tactical size and weight. The successful implementation of the theory will allow a compression ratio of about 500:1. Moreover , it is now possible to produce good quality synthetic speech from a digital signal representing the phonemic content of a message, so that the development of a complete system awaits the measurement of the acoustical correlates of the unspecified features.

In addition to providing an up-to-date synopsis of speech compression methods, this paper includes the design of an acoustical input device, compatible to the CDC 1604, via the CDC 160. This device will provide a means for investigating the "method of distinctive features" as well as any other method related to speech recognition and/or communication. Finally, computer flow-graphs are given which will aid future engineers in initiating research of the method at the United States Naval Postgraduate School.

16. Bibliography

1. R. Jakobson, C. G. M. Fant and M. Halle, Preliminaries to speech analysis, Tech. Report No. 13, Acoustics Lab., MIT, 1952.

2. K. N. Stevens, Toward a model for speech recognition, J. Acoust. Soc. Am., 32,pp. 47-55, 1960.

3. C. Shannon, Prediction and entropy of printed English, Bell System Tech. J., 30, pp. 50-64, 1951.

4. J. R. Pierce, J. E. Karlin, Information rate of a human channel, Proc. IRE, 45, p. 368, 1957.

5. J. L. Flanagan, A resonance-vocoder and baseband complement: A hybrid system for speech transmission, IRE Trans. on Audio, v. AU-8, No. 3, May-June, 1960.

6. K. N. Stevens, M. H. L. Hecker and K. D. Kryter, An evaluation of speech compression systems, Tech. Doc. Report No. RADC-TDR-62-171, USAF Contract No. AF 30(602)-2235, Proj. No. 4519, Task No. 45350, Mar., 1962.

7. M. R. Schroeder, Vocoders for military use, Proc. of Sem. on Speech Compression and Processing, 1, Sept., 1959.

8. H. Dudley, The carrier nature of speech, Bell System Tech. J., 19, 1940.

9. H. Fletcher, Speech and hearing in communications, D. Van Nostrand Co., 1955.

10. F. H. Slaymaker, Characteristics of modern vocoders and remaining problems, Proc. of Sem. on Speech Compression and Processing, 1, Sept. 1959.

11. C. P. Smith, An approach to speech bandwidth compression. Proc. of Sem. on Speech Compression and Processing, 2, Sept., 1959.

12. S. E. Gerber and W. H. Morris, Intelligibility study of the Hughes Digital Vocoder, J. Acoust. Soc. Am., 33, p. 835A, 1961.

13. M. R. Schroeder, Correlation techniques for speech bandwidth compression, Paper presented at the 12th Annual meeting of the Audio Engineering Society, Oct., 1960.

14. N. Wiener, Generalized harmonic analysis, Acta Math. 55, p. 117, 1930.

15. R. M. Fano, Short-time autocorrelation functions and power spectra, J. Acoust. Soc. Am., 22, p. 546, 1950.

16. R. Biddulph, Short-term autocorrelation analysis and correlatograms of spoken digits, J. Acoust. Soc. Am., 26, p. 539, 1954.

17. M. R. Schroeder and B. S. Atal, Generalized short-time power spectra and autocorrelation functions, J. Acoust. Soc. Am., 34, p. 1679, 1962.

18. M. R. Schroeder, T. H. Crystal, An autocorrelation vocoder, Master's Thesis, MIT, June, 1960.

19. C. G. M. Fant, Transmission properties of the vocal tract with application to the acoustic specification of phonemes, Tech. Report No. 12, Acoustics Lab., MIT, Jan., 1952.

20. J. L. Flanagan, Development and testing of a formant-coding speech compression system, J. Acoust. Soc. Am., 28, pp. 1099-1106. Nov., 1956.

21. J. L. Flanagan, Automatic extraction of formant frequencies from continuous speech, J. Acoust. Soc. Am., 28, pp. 110-118, 1956.

22. J. L. Flanagan, Note on the design of "terminal-analog" speech synthesizers, J. Acoust. Soc. Am., Feb., 1957.

23. K. N. Stevens, Synthesis of speech by electrical analog devices, J. Audio Eng. Soc., 4, pp. 2-8, 1956.

24. F. S. Cooper, Basic factors in speech perception and applications to speech processing, Proc. of Sem. on Speech Compression and Processing, 1, Sept., 1959.

25. G. W. Hughes, The recognition of speech by machine, Tech. Report No. 395, Research Laboratory of Electronics, MIT, May, 1961.

26. C. G. M. Fant, Acoustic theory of speech production, The Hague: Mouton and Co., 1960.

27. M. Halle, The sound pattern of Russian, The Hague: Mouton and Co., 1959.

28. Hughes Communications Division, Speech recognition studies, Exhibit B., 1952.4/289/A1205-001, Feb., 1963.

29. Acoustics Lab, MIT, Speech compression research, USAF Cont. No. AF 19(604)-626, Feb., 1957.

30. Hughes Communications Division, Speech transmission evaluation, Exhibit B, 1952.1/55/A0700-002, Nov., 1962.

31. M. Halle, G. W. Hughes, and J. Radley, Acoustic properties of stop consonants, J. Acoust. Soc. Am., 29, pp. 107-116, 1957.

32. G. Peterson and H. Barney, Control methods used in a study of the vowels, J. Acoust. Soc. Am., 24, pp. 175-184, 1952.

33. J. D. Hollabaugh, Construction and partial evaluation of a vocoder transmitter, TR-DCL-62-14, USNPS, Dec., 1962.

34. P. Denes, Computer processing of acoustic and linguistic information in automatic speech recognition, Tech. Report No. AF61(514)-1176, Univ. College, London, England, Mar., 1962.

35. H. M. Kaplan, Anatomy and physiology of speech, McGraw-Hill Book Co., Inc., p. 128, 1960.

36. W. C. Michels, Sr. Ed., The international dictionary of physics and electronics, D. Van Nostrand Co., Inc., p. 623, 1956.

37. J. L. Flanagan, A speech analyzer for a formant-coding compression system, Scient. Report No. 4, USAF Contract No. AF 19(604)-626, Acoustics Lab, MIT, May, 1955.

APPENDIX A

17. Definitions

1. formant frequency is interpreted fundamentally as the frequency of a normal mode of vibration of the human vocal mechanism. As such, formant frequency is a complex number. During the utterance of certain sounds, notably the vowels, the normal modes of vibration of the vocal system are manifested in the acoustic output of the speaker as maxima in the spectra of these sounds. The frequencies of these spectral maxima (called "formants"), are closely related to the complex numbers representing the normal modes of vibration of the vocal tract. The term formant frequency, therefore, is applied both to the normal modes (or natural frequencies) of the vocal system and to their manifestations in the speaker's acoustic output, i. e. the frequencies of the spectral maxima /19/.

2. glottis (rima glottidis) is the variable opening between the vocal folds. It is the narrowest part of the laryngeal cavity /35/.

3. Nyquist rate, signaling. In transmission, if the essential frequency range is limited to B cycles per second, 2B is the maximum number of code elements per second that can be unambiguously resolved, assuming the peak interference is less than half a quantum step. This rate is generally referred to as signaling at the Nyquist

rate, and ½B is called the Hyquist interval /36/.

4. intelligibility is a measure of the ability of the listener to reproduce what was said. An intelligibility score of 80-85% on standard Harvard PB word lists implies about 90-95% sentence intelligibility.

5. speaker recognition is a measure of the ability of the listener to identify the test speaker, provided the test speech is clearly audible. Additionally, this may also be a measure of the ability of the listener to differentiate between two or more test speakers.

6. naturalness is a measure of how "true" the test speech is. For example, a human can tell the difference among flute, piano and clarinet even if they all play the same tone. In addition, moods and emotions, evidenced in the voice, should be clearly apparent in the test speech. Technically, this is a measure of how well the transmitting equipment maintains the relative strengths of the test speech's overtones, the harmonics of the fundamental pitch.

7. quality is a measure of the fidelity and is concerned with how faithfully voice naturalness, as well as intelligibility, is preserved. For example, you might recognize a test voice as belonging to your friend Joe, but due to noise, the "quality" of the reception is poor.

APPENDIX A

(Continued)

8. glide is a term used to describe the sound emitted when the vocal tract has assumed a position to emit one vowel and instantly changes and emits another. For example, the word "we".
9. semi-vowels consist of the nasals and glides whose spectrums contain a formant structure similar to that of vowels.
10. General American English is the dialect spoken by most cultured speakers of American English.

APPENDIX B

18. The Phonemes of English

1. /i/ as in beet, green.
2. /I/ as in sit, in.
3. /u/ as in pool.
4. /U/ as in pull.
5. /ɛ/ as in set, get.
6. /e/ as in chaotic, debris.
7. /æ/ as in sat.
8. /o/ as in notation.
9. /ɔ/ as in all, horse.
10. /a/ as in hot.
11. /ʌ/ as in sun, up.
12. /r/ as in rear.
13. /l/ as in lily.
14. /w/ as in witch.
15. /m/ as in beam.
16. /j/ as in yet.
17. /n/ as in dine.
18. /ŋ/ as in ring.
19. /f/ as in foul.
20. /s/ as in song.
21. /ʃ/ as in she.
22. /θ/ as in thin.
23. /ð/ as in that.
24. /v/ as in vile.
25. /z/ as in buzz.
26. /ʒ/ as in casual.
27. /dʒ/ as in jumbo.
28. /tʃ/ as in child.
29. /p/ as in pit.
30. /t/ as in too.
31. /k/ as in kind.
32. /b/ as in bit.
33. /d/ as in done.
34. /g/ as in go.
35. /h/ as in him.

APPENDIX C

19. Filter Specifications

CHANNEL NUMBER	BAND LIMITS	CENTER FREQUENCY	BANDWIDTH	WEIGHTING
1	115-165	140	50	0.386
2	165-215	190	50	0.450
3	215-265	240	50	0.505
4	265-315	290	50	0.555
5	315-365	340	50	0.602
6	365-415	390	50	0.644
7	415-465	440	50	0.685
8	465-520	493	55	0.725
9	520-580	550	60	0.765
10	580-645	613	65	0.808
11	645-720	683	75	0.853
12	720-805	763	85	0.900
13	805-900	853	95	0.950
14	900-1005	953	105	1.00
15	1005-1120	1063	115	1.00
16	1120-1255	1188	135	1.10
17	1255-1410	1333	155	1.20
18	1410-1585	1498	175	1.30
19	1585-1780	1683	195	1.40
20	1780-2005	1893	225	1.50
21	2005-2280	2143	275	1.38
22	2280-2580	2430	300	1.25

APPENDIX C

(Continued)

CHANNEL NUMBER	BAND LIMITS	CENTER FREQUENCY	BANDWIDTH	WEIGHTING
23	2580-2930	2755	350	1.12
24	2930-3280	3105	350	1.00
25	3280-3680	3480	400	1.00
26	3680-4180	3930	500	1.00
27	4180-4780	4480	600	1.00
28	4780-5480	5130	700	1.00
29	5480-6280	5880	800	1.00
30	6280-7280	6780	1000	1.00
31	7280-8500	7890	1220	1.00
32	8500-10000	9250	1500	1.00